

AD-A267 162



2

NAVAL POSTGRADUATE SCHOOL Monterey, California

S DTIC
ELECTE
JUL 28 1993
A **D**



THESIS

THE EFFICACY OF MACHINE LEARNING PROGRAMS FOR
NAVY MANPOWER ANALYSIS

by

Dennis E. Pytel, Jr.

March 1993

Principal Co-Advisor:
Principal Co-Advisor:

George W. Thomas
Daniel R. Dolk

Approved for public release; distribution is unlimited.

93-16797



Unclassified

Security Classification of this page

REPORT DOCUMENTATION PAGE				
1a Report Security Classification: Unclassified		1b Restrictive Markings		
2a Security Classification Authority		3 Distribution/Availability of Report		
2b Declassification/Downgrading Schedule		Approved for public release; distribution is unlimited.		
4 Performing Organization Report Number(s)		5 Monitoring Organization Report Number(s)		
6a Name of Performing Organization Naval Postgraduate School	6b Office Symbol (if applicable) 36	7a Name of Monitoring Organization Naval Postgraduate School		
6c Address (city, state, and ZIP code) Monterey CA 93943-5002		7b Address (city, state, and ZIP code) Monterey CA 93943-5002		
8a Name of Funding/Sponsoring Organization	6b Office Symbol (if applicable)	9 Procurement Instrument Identification Number		
Address (city, state, and ZIP code)		10 Source of Funding Numbers		
		Program Element No	Project No	Task No
		Work Unit Accession No		
11 Title (include security classification) The Efficacy of Machine Learning Programs for Navy Manpower Analysis				
12 Personal Author(s) Dennis E. Pytel, Jr.				
13a Type of Report Master's Thesis	13b Time Covered From To	14 Date of Report (year, month, day) 1993, March, 25	15 Page Count 101	
16 Supplementary Notation The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.				
17 Cosati Codes		18 Subject Terms (continue on reverse if necessary and identify by block number)		
Field	Group	Subgroup		
19 Abstract (continue on reverse if necessary and identify by block number) This thesis investigated the efficacy of two machine learning programs for Navy manpower analysis. Two machine learning programs, AIM and IXL, were compared to conventional statistical techniques. A large manpower data set and a logistic regression equation were obtained. The same data set was used to generate models from the two commercial machine learning programs. Using a held out sub-set of the data the capabilities of the three models were evaluated. AIM generated results comparable to those of the logistic regression equation; both in number of correct predictions and computed partial effects of the independent variables. IXL had significantly fewer correct predictions than the other two models and does not support evaluation of partial effects. The author recommended further investigation of AIM's capabilities, and testing in an operational environment.				
20 Distribution/Availability of Abstract <input checked="" type="checkbox"/> unclassified/unlimited <input type="checkbox"/> same as report <input type="checkbox"/> DTIC users		21 Abstract Security Classification Unclassified		
22a Name of Responsible Individual George W. Thomas		22b Telephone (include Area Code) (408) 656-2741	22c Office Symbol AS/TE	

DD FORM 1473,84 MAR

83 APR edition may be used until exhausted

security classification of this page

All other editions are obsolete

Unclassified

Approved for public release; distribution is unlimited.

The Efficacy of Machine Learning Programs For Navy Manpower
Analysis

by

Dennis Eric Pytel, Jr.
Lieutenant, United States Navy
A.B, University of California, Davis, 1985

Submitted in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT

from the

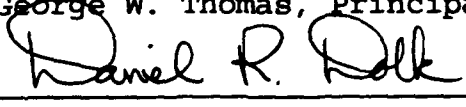
NAVAL POSTGRADUATE SCHOOL
March 1993

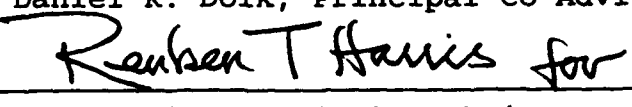
Author:


Dennis E. Pytel, Jr.

Approved by:


George W. Thomas, Principal Co-Advisor


Daniel R. Dolk, Principal Co-Advisor


David R. Whipple, Chairman
Department of Administrative Sciences

ABSTRACT

This thesis investigated the efficacy of two machine learning programs for Navy manpower analysis. Two machine learning programs, AIM and IXL, were compared to conventional statistical techniques. A large manpower data set and a logistic regression equation were obtained. The same data set was used to generate models from the two commercial machine learning programs. Using a held out sub-set of the data the capabilities of the three models were evaluated. AIM generated results comparable to those of the logistic regression equation; both in number of correct predictions and computed partial effects of the independent variables. IXL had significantly fewer correct predictions than the other two models and does not support evaluation of partial effects. The author recommended further investigation of AIM's capabilities, and testing in an operational environment.

DTIC QUALITY INSPECTED 5

Accession For	
NTIS	<input checked="" type="checkbox"/>
CRA&I	<input type="checkbox"/>
DTIC	<input type="checkbox"/>
TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and/or Special
A-1	

TABLE OF CONTENTS

I. INTRODUCTION	1
A. BACKGROUND	1
B. THESIS OBJECTIVES	3
C. RESEARCH QUESTIONS	3
D. ORGANIZATION OF THE STUDY	3
II. LITERATURE REVIEW AND CONCEPT LEARNING	5
A. INTRODUCTION	5
B. LITERATURE REVIEW	6
C. CONCEPT LEARNING	8
D. ABDUCTORY INDUCTION MECHANISM (AIM)	10
1. Learning Technique	10
2. Output Format	14
E. INDUCTION ON EXTREMELY LARGE DATABASES (IXL)	16
1. Learning Technique	16
2. Output Format	23
III. DATA AND METHODOLOGY	27
A. INTRODUCTION	27
B. DATA	28
1. Data Source	28
2. Thesis Data Set	29

3. Variable Definitions	31
a. Dependent Variable (STATUS)	31
b. Independent Variables	31
(1) Demographic Variables	31
(2) Military Characteristics	33
(3) Educational Accomplishment	34
(4) Satisfaction with Military Lifestyle and Benefits	35
C. METHODOLOGY	37
1. Multivariate Logistic Regression Analysis	37
2. AIM Model	38
3. IXL Rules	39
D. EXPANDED VARIABLE DATA SET	41
IV. RESULTS	44
A. LOGISTIC REGRESSION	44
B. AIM	45
C. IXL	46
D. EXPANDED VARIABLE AIM MODEL	47
V. DISCUSSION	48
A. INTRODUCTION	48
B. PREDICTION	49
1. Overall Comparison	49
2. Best AIM Network	50
C. EFFECT OF CHANGES IN INDEPENDENT VARIABLES	50

1. Introduction	50
2. Computed Effects	52
3. Comparison	55
D. EXPANDED VARIABLE AIM NETWORK	57
E. OTHER STRENGTHS AND WEAKNESSES	60
1. Documentation	60
a. AIM	60
b. IXL	61
2. Output Interpretation	62
a. AIM	62
b. IXL	63
3. Model and Variable Significance	63
VI. CONCLUSIONS AND RECOMMENDATIONS	65
A. CONCLUSIONS	65
B. RECOMMENDATIONS	66
APPENDIX A. LOGISTIC REGRESSION EQUATION	67
APPENDIX B. AIM NETWORKS	68
APPENDIX C. IXL RULES	71
A. NUMBER OF CORRECT PREDICTIONS BY RULE	71
B. IXL RULES	72
APPENDIX D	86

LIST OF REFERENCES	88
------------------------------	----

INITIAL DISTRIBUTION LIST	90
-------------------------------------	----

LIST OF FIGURES

Figure 1: Four Input, Three Layer Polynomial Network. .	14
Figure 2: Lineage of IXL	16
Figure 3: One level decision tree.	18
Figure 4: Two level decision tree.	19
Figure 5: Decision tree with class names.	19
Figure 6: Generalization tree of descriptive terms. . .	21
Figure 7: Example IXL rule.	24
Figure 8: Sample IXL rule with two identifying concepts.	25
Figure 9: Default Network (CPM = 1 and number of levels = 4)	68
Figure 10: Best AIM Network (CPM = .5 and number of levels = 4)	69
Figure 11: Expanded Variable AIM Network	86

LIST OF TABLES

Table I 1985 DOD SURVEY OF OFFICER AND ENLISTED PERSONNEL:

TOPIC AREAS	29
Table II ROTATED FACTOR PATTERN SCORES	36
Table III VALUES FOR NEW MILLIFE AND MILBENE VARIABLE .	40
Table IV NUMBER OF CORRECT PREDICTIONS FOR AIM NETWORKS	45
Table V CHANGE IN PROBABILITY OF ENLISTMENT FROM BASE CASE.	54
Table VI CHANGE IN PROBABILITY OF ENLISTMENT FROM BASE CASE: EXPANDED VARIABLE NETWORK.	58
Table VII ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATORS. .	67

I. INTRODUCTION

A. BACKGROUND

Machine learning is a subfield of artificial intelligence. Machine learning generally refers to the ability of a program to discover or learn information by itself, given a set of examples. Machine learning is not new, but interest in the field has increased significantly as new formal methods and implementation techniques have been developed. Advances in computer technology have also affected machine learning. High-speed, inexpensive microcomputers have given individuals working at their desks processing power which, in the past, was only available to those with access to mainframe computers.

Because the average individual works only eight hours during a day, the proliferation of microcomputers has created an enormous amount of computing capability which is unused. Another important change created by computer advances is that organizations are collecting more data which is stored in relational databases. Often data is collected without clear understanding of how the information can be useful to the organization. Organizations searching for ways to use their data and excess processing capabilities to gain competitive advantage, led to the introduction of so-called "data mining"

programs. These software packages use machine learning techniques to discover unknown, or unexplored, relationships between variables in a database which may be useful to the organization. An important characteristic of these programs is that they are designed to be used by people with little or no knowledge of formal research techniques. This characteristic makes them potentially useful to a large number of users, including Navy manpower planners.

Navy manpower planners have an enormous amount of data and processing capability available to them. However, experienced researchers are consistently in short supply. Normally the Navy contracts with outside organizations to provide research services. There are two significant problems with using contract services: cost and time. With Navy budgets decreasing, the cost of contracting becomes increasingly sensitive. The contracting process also takes time, which makes it difficult to use when quick answers are needed. Machine learning programs may be a partial solution to both of these problems. If these programs are capable of providing timely and accurate information, and are relatively easy to use, they could provide useful answers to planners in some cases, and help direct future research in other cases. Decreased reliance on outside research would save the Navy money, possibly in amounts orders of magnitude larger than the respective procurement costs of the software packages. The

potential savings justify a closer examination of machine learning programs and their effectiveness.

B. THESIS OBJECTIVES

The objective of this thesis is to study the efficacy of personal computer (PC) based machine learning programs for Navy manpower analysis. Specifically, I will assess the capabilities and performance of two commercial machine learning programs: AIM, produced by AbTech Corporation; and IXL, produced by IntelligenceWare Inc.

C. RESEARCH QUESTIONS

This thesis will attempt to answer the following questions:

- Can machine learning programs such as "IXL" and "AIM" enhance Navy manpower analysis?
- What are the strengths and weaknesses of machine learning programs?
- Do these different programs, when run on the same data set, generate comparable results?
- How do results generated by machine learning programs compare with equations generated by conventional regression techniques?

D. ORGANIZATION OF THE STUDY

The first phase of research will examine the most relevant subfield of machine learning, concept learning, and specifically, how the two programs examined in this thesis generate results. The programs will then be tested to

determine which program best predicts reenlistment of U.S. Navy enlisted personnel. I will also develop a model using conventional regression techniques. Both programs will use an identical data set generated from selected observations and variables taken from the 1985 *Department of Defense Survey of Officer and Enlisted Personnel*. Comparisons will be made between results obtained using the machine learning programs and the regression model. The final portion of the thesis will be an assessment of the usefulness and accuracy of machine learning programs for manpower analysis.

II. LITERATURE REVIEW AND CONCEPT LEARNING

A. INTRODUCTION

This chapter will examine the most important subfield of machine learning related data-mining programs, concept learning. The first section will examine previous research on machine learning and related areas. The second section will define and examine some general principles of concept learning. The final two sections will examine how each of the machine learning programs selected for this thesis function. The two programs are commercially produced and distributed and, for proprietary reasons, the publishers have not released the exact algorithms that their software uses. However, the publishers do give a general overview of how the programs work, as well as identifying previous work that influenced development of their software. For each package, this chapter will:

1. describe the learning technique the software utilizes and,

2. provide a sample output and interpret the results.

The final section will examine available literature on machine learning and machine learning techniques.

B. LITERATURE REVIEW

Little research has been done specifically on either AIM or IXL. AIM was found to provide accurate answers in less time than standard neural network techniques. According to a high-technology update by the Strategic Defense Initiative Organization, "Studies performed by the U.S. Air Force's Space Systems Division found that AIM developed networks in less time and more than 100 times more accurately than neural networks. Lockheed uses AIM in their Pilots Associate program. The system acts as an electronic crew member by analyzing air combat situations and performing less critical tasks for the pilot. Lockheed found AIM to do a more accurate job than traditional psychological techniques (80% accuracy as compared to 30%). In a software review published in *AI Expert*, Angell and Murphy found AIM to be faster and easier to use than traditional neural networks. [Ref.1, p.50]

AIM is often compared favorably to neural networks for solving a wide range of problems. Therefore, it is relevant to examine how well neural networks perform as compared to other techniques.

Sands used computer simulated personnel data to compare ordinary least-squares (OLS) linear regression and neural networks. His study examined different functional forms (linear and curvilinear), sample sizes (100, 500, and 5000), errors (deviation from ideal functional form), and sample splits (proportions of sample used for network training versus

network evaluation). He found that the predictive capabilities of OLS regression and neural networks were not significantly different if the underlying functional form of the data was linear. However, neural networks performed significantly better than OLS regression if the functional form was not linear. He also found that neural networks performed particularly well using large samples. [Ref.2, p. 21]

Wiggins compared neural networks and regression and found that a model developed using a neural network was significantly more accurate than an OLS model for predicting enlisted personnel performance on the U.S. Air Force's walk through performance test [Ref.3, p.11]. Marquez found that neural network models perform best under conditions of high noise and low sample size. With less noise and larger sample size they are less competitive [Ref.4, p.10]. Hill, O'Connor, and Remus evaluated time series forecasting and found neural networks to perform as well or better than classical forecasting models [Ref.5, p.17].

Weiss and Kapouleas compared statistical pattern recognition techniques, neural networks and machine learning classification methods similar to IXL. They found that machine learning methods were at least as effective as statistics or neural networks in most cases [Ref.6, p. 182].

Mooney compared results obtained using ID3 and two methods of training neural networks. He found that ID3 ran

significantly faster than neural networks, and the probability of correct classification was about the same. He also found that neural networks trained using back-propagation (an error correcting technique) were more accurate if the data set was noisy. [Ref.7, p.174]

In summary, while not definitive, it appears that machine learning programs are at least as accurate as other methods of concept learning and pattern recognition.

C. CONCEPT LEARNING

Concept learning is a subfield of machine learning. There are two categories of concept learning: knowledge acquisition and skill refinement. This chapter will concentrate on the knowledge acquisition category, the goal of data mining programs. The goal of concept learning is to extract the important features which describe all members of a concept. AIM and IXL both learn concepts using *induction* or *inductive inference*, which consists of extracting general rules, concepts, or other data structures from specific facts. Induction differs from deduction because, although one can logically infer facts from the generalization obtained via induction, one cannot, in general, deduce the generalization from the facts using the strict rules of logical inference. Therefore, the inference from the specific facts to the generalization is not truth-preserving. It is, however, falsity-preserving. For example:

Suppose we have facts *F* and hypothesis *H*. If the inference used to derive *H* from *F* is deduction, then if *F* is true, *H* must necessarily be true. If the inference used to derive *H* from *F* is induction, then if *F* is true, *H* may or may not be true. However, if *H* is inductively inferred from *F*, then if *H* is true, *F* must be true. Furthermore, if some facts falsify *F*, then they must also falsify *H*, i.e., induction is falsity-preserving. [Ref.8, p. 2]

Researchers using these software packages need to be cognizant of this limitation of induction. All results generated by these software packages need to be examined carefully to ensure that they are an accurate representation of the world and not peculiar to the particular examples being used.

Another important distinction in concept learning is whether the program is capable of *supervised* and/or *unsupervised* learning. In supervised learning the system learns from a set of known correctly classified cases, that "supervises" the choice of learning cases. Unsupervised learning, or *clustering*, uses data where no classifications are given. The goal of unsupervised learning is to identify clusters of patterns which are similar, thus identifying potential classes. Both AIM and IXL perform supervised learning, but only IXL can perform unsupervised learning.

With the concepts of induction and supervised and unsupervised learning as a background, the next sections of this chapter will examine how AIM and IXL generate results. For each package I will determine how the program learns concepts and how it presents the results to the user.

D. ABDUCTORY INDUCTION MECHANISM (AIM)

1. Learning Technique

AIM is a numeric modeling tool which, given a database of examples, automatically synthesizes a mathematical model of the relationships in the data. AIM generates a network of functional elements called a *polynomial network*. According to AbTech, the power of polynomial networks is derived from the ability to deal with complex problems by subdividing them into smaller, simpler ones. Networks simplify induction because only the relationships among small subsets of variables need to be solved at any given time. [Ref.9, p. 2-4] AbTech calls the general processes that AIM uses *network regression*TM.

Network regression combines the network concept from *neural networks* and advanced regression techniques to create a polynomial network. A polynomial network is a network of functional nodes. Each node contains a mathematical function which computes an output given a number of inputs. The final network is a layered network of feed-forward functional elements. Feed-forward elements use the output from the one layer and original input variables as inputs to the next layer. Information flows from the input variables through the network to the output variables.

AIM automatically determines a "best" network structure by minimizing a modeling criterion called predicted

square error (PSE). The predicted square error is given by:

$$PSE = FSE + KP$$

where FSE is the fitted square error of the model on the training data and KP is a complexity penalty. The complexity penalty is determined by AIM using the equation [Ref.10, p.2-7]:

$$KP = CPM * \frac{2K}{N} * s_p^2$$

where K is the total number of coefficients, N is the number of training data observations or cases, and s_p^2 is an a-priori estimate of the true unknown model variance. As N increases or s_p^2 decreases, more complexity is allowed. CPM is the Complexity Penalty Multiplier, which is set by the user prior to model synthesis. A higher value for CPM increases the impact of the complexity penalty term, which will result in a simpler network. If, for example, the user increases the CPM from 1 (the default value) to 2, the value of the complexity penalty doubles. In order to minimize PSE, AIM will perform a tradeoff between FSE and the KP. The only variable that AIM controls in the KP equation is K (the total number of coefficients in the network). Therefore, to offset the increase in the CPM, AIM will create a simpler network with fewer coefficients. Using PSE allows AIM to perform a tradeoff between model complexity and accuracy to generate the best possible model without overfitting the data. Overfitting occurs when the model becomes so specific to the training data

that it does a poor job of describing new data. The assumption is that simpler models are more general and superior for describing future data. Using PSE allows AIM to synthesize networks with little user intervention. Because it is possible to generate a model with little user interaction, the user is not required to possess the specialized knowledge that using either neural networks or regression techniques require.

AbTech's use of the term *network regression* invites comparisons between AIM and both neural networks and regression techniques. AIM primarily uses networks in order to subdivide complex problems into simpler ones. Compared to neural networks, AIM has fewer, more powerful nodes. A neural network node gives an output using a weighted sum of inputs. An AIM node can use polynomial equations of degree three, and can also use interaction terms between input variables. AIM also uses statistical methods and a modeling criterion to select the network structure automatically, whereas neural networks require the user to select the network synthesis technique, structure, and numerous parameters prior to inputting data. Parameters and structure are then modified using trial and error to obtain the best results. This allows AIM to generate results faster and more accurately than neural networks in many cases. One significant advantage of neural networks is they are capable of unsupervised learning, where inputs and corresponding outputs are not known.

AIM network synthesis can be classified as a form of non-parametric regression. The primary limitation of regression is that in order to generate accurate results it is necessary to know the underlying form of the relationships. Researchers cannot assume a general polynomial equation and determine the coefficients using multiple linear regression because the number of independent coefficients grows factorially as a function of the number of variables and degrees. For example, a 9-variable, 27-degree complete polynomial requires 94,143,280 different coefficients to be determined. AIM can approximate a large number of these functions using a three layer network with only 104 independent coefficients [ibid., p.2-15]. If the underlying functional form is known, using regression is more appropriate than AIM. According to AbTech, the primary advantages of AIM over other non-parametric techniques is that it produces very compact and rapidly executable models, gives a practical method for applying non-parametric regression, and can be used effectively by people who do not have any knowledge of advanced statistical theory [ibid., p.2-14].

2. Output Format

AIM displays the results of the network in graphical form. An example is shown in Figure 1.

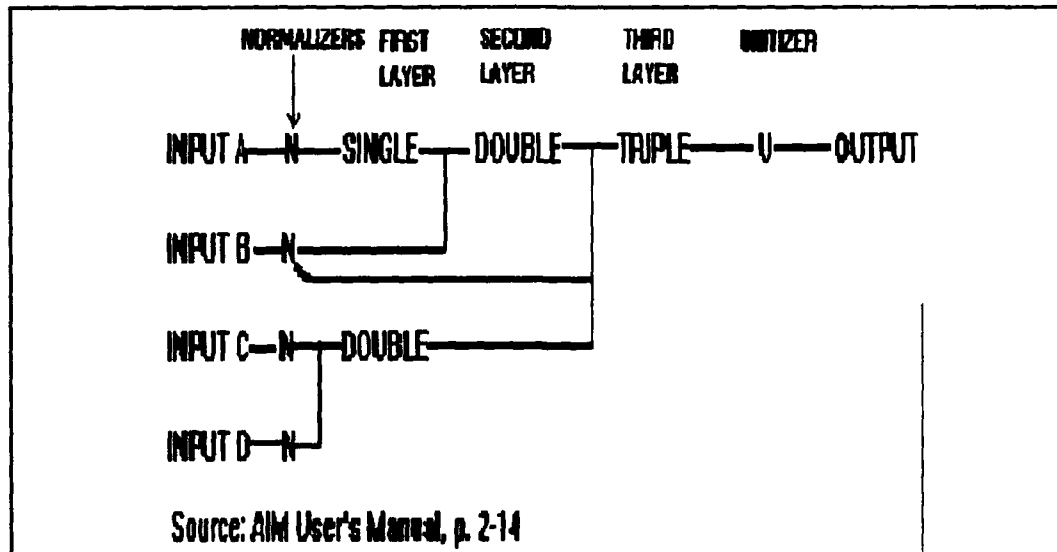


Figure 1: Four Input, Three Layer Polynomial Network.

AIM uses seven types of nodes. The algebraic form of each element is shown in the equations below where w_n are the coefficients determined by AIM and x_n are the input variables. All of the terms in an equation may not appear in a node since AIM will throw out terms which do not contribute significantly to the solution.

1. Singles: $w_0 + (w_1x_1) + (w_2x_1^2) + (w_3x_1^3)$
2. Doubles: $w_0 + (w_1x_1) + (w_2x_2) + (w_3x_1^2) + (w_4x_2^2) + (w_5x_1x_2)$
 $+ (w_6x_1^3) + (w_7x_2^3)$

$$3. \text{ Triples: } w_0 + (w_1x_1) + (w_2x_2) + (w_3x_3) + (w_4x_1^2) + (w_5x_2^2) + (w_6x_3^2) \\ + (w_7x_1x_2) + (w_8x_1x_3) + (w_9x_2x_3) + (w_{10}x_1x_2x_3) + (w_{11}x_1^3) \\ + (w_{12}x_2^3) + (w_{13}x_3^3)$$

Singles, Doubles, and Triples are names based upon the number of input variables. Note that these elements are third degree polynomial equations and that doubles and triples have cross terms.

$$4. \text{ White: } w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_nx_n$$

The white element consists of the linear weighted sums of all the outputs of the previous layer.

$$5. \text{ Normalizers: } w_0 + (w_1x_1)$$

Normalizers transform all of the original input variables into a relatively common region with a mean of zero and a variance of one using mean-standard deviation normalization.

$$6. \text{ Unitizers: } w_0 + (w_1x_1)$$

A unitizer converts the range of the network outputs to a range with a mean and variance of the output values used to train the network. This takes place at the end of the network, and essentially reverses the effects of Normalizers.

7. Wire: The wire element is used for a network that consists of only a normalizer and a unitizer. [ibid., p. 2-5]

Once the network has been synthesized, each node can be individually examined to determine coefficient values. Once the user is satisfied with the network there are two ways to use it. The first is the Query function that allows the user to give the network input values which are used to compute output values. AIM also generates generic "C" language computer source code which can be integrated into an application program.

E. INDUCTION ON EXTREMELY LARGE DATABASES (IXL)

1. Learning Technique

IXL combines machine learning and statistical techniques in order to discover "logical relationships". The relationships are reported as a series of rules rather than equations. The advantage of using rules is that they are more readable than equations and do not need interpretation by a person knowledgeable in statistics or mathematics. Figure 2 shows how some previous machine learning programs are related to IXL.

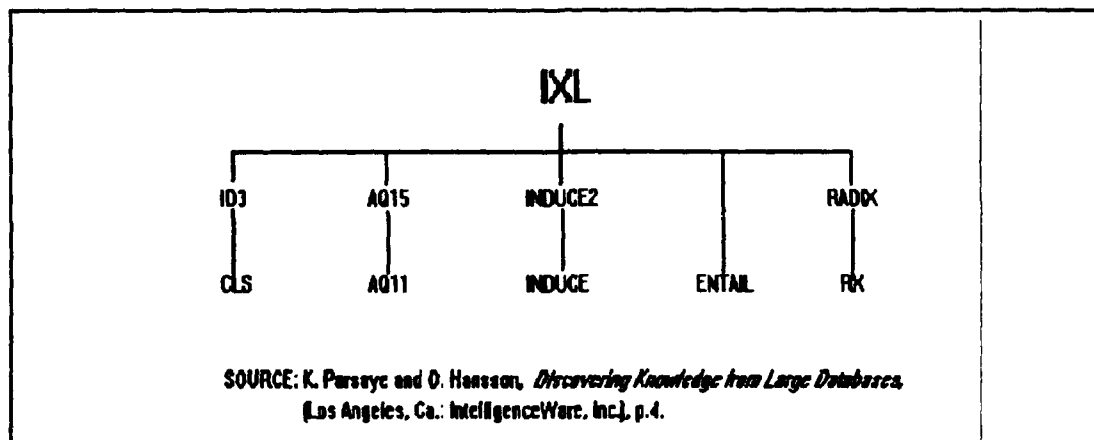


Figure 2: Lineage of IXL

The most relevant of these programs are ID3 and AQ15. Each of these programs is discussed below.

ID3 takes a set of examples about some problem and induces a decision tree or set of rules that captures the decision-making knowledge about the problem. The ID3

algorithm is a descendant of Hunt's Concept Learning System (CLS). CLS solves single-concept learning tasks and uses the learned concepts to classify new examples. CLS can discover a decision tree for a collection of examples and use this tree to classify a new example into one of the two classes. An example is classified by starting at the root of the tree, making tests, and following the branches until a node is reached, which indicates whether the example is part of the class indicated. Unlike CLS, ID3 can work with subsets of the examples in order to solve more complex problems. The ID3 algorithm follows 4 major steps [Ref.11, p.51]:

1. Select a random subset of size W from the entire set of training examples (W is called the window).
2. Apply the CLS algorithm to form the decision tree or rule for the window.
3. Scan the entire set of examples (not just the window) to find exceptions.
4. If there are exceptions, insert some of them into the window and repeat step 2; otherwise stop and display the latest rule.

This algorithm iteratively converges on a rule that captures the concept. The process continues until either all of the examples are of the same class (a leaf) or the number of remaining examples falls below some minimum value. The eventual outcome is a tree in which each leaf carries a class name, and each interior node specifies an attribute which must be tested with a branch corresponding to each possible value of that attribute. To illustrate this process, consider

collection "C" below. Each object in C is described in terms of three attributes: "height" {tall,short}, "hair" {dark, red, blond}, and "eyes" {blue, brown} and is followed by a '+' or '-' to indicate the class to which it belongs.

C= short,blond,blue:+short,dark,blue:-tall,dark,brown:-
tall,blond,brown:-tall,dark,blue:-short,blond,brown:-
tall,red,blue:+tall,blond,blue:+

If the second attribute is used to form the root of the decision tree, this yields the tree shown in Figure 3. The subcollections corresponding to 'dark' and 'red' contain objects of only one class and do not require further work. If we use the third attribute to test for the 'blond' branch, this yields the tree in Figure 4. Now all of the subcollection contain only one class so we can replace each subcollection by the class name to yield the tree in Figure 5.

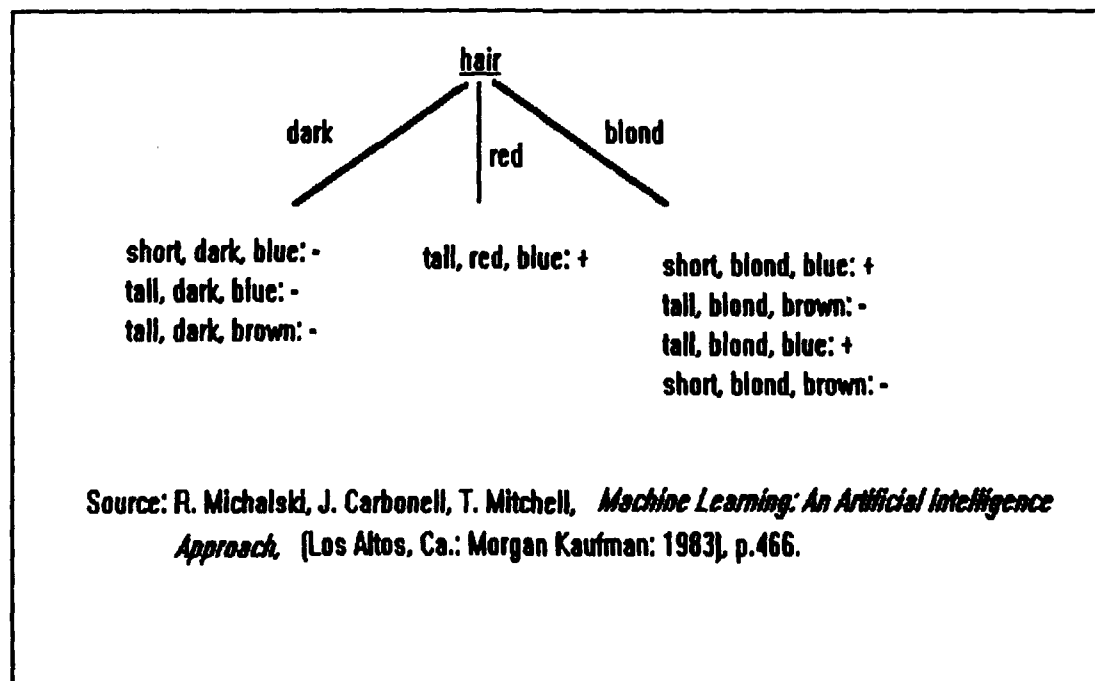


Figure 3: One level decision tree.

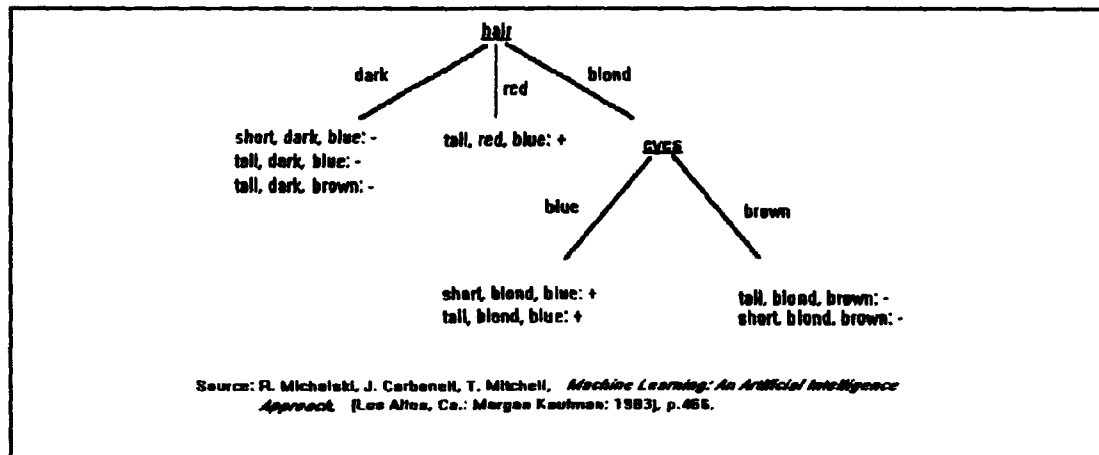


Figure 4: Two level decision tree.

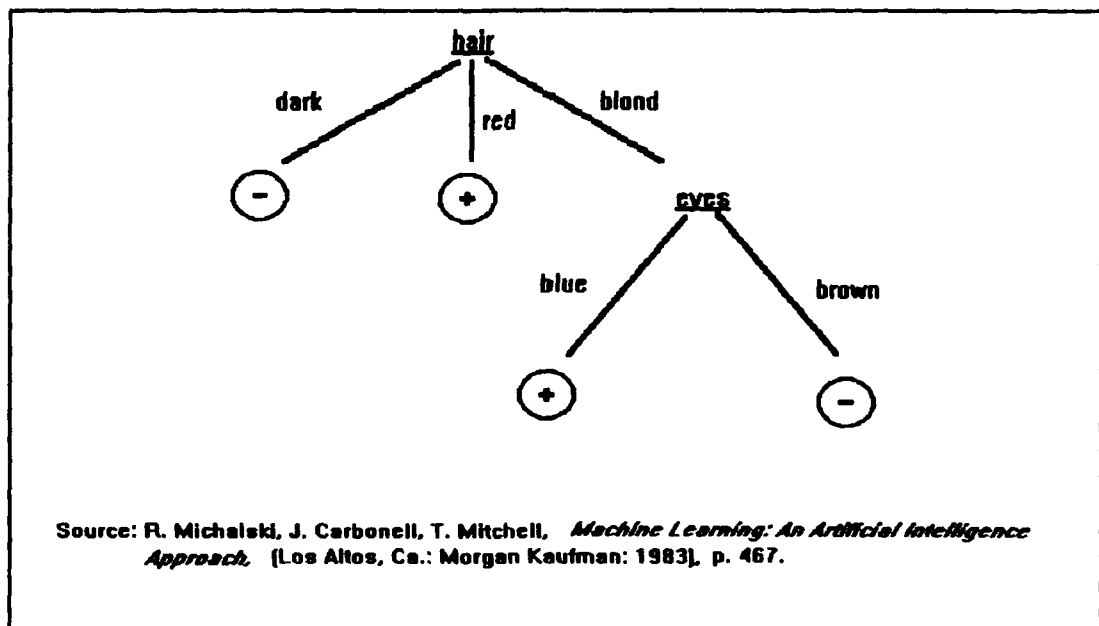


Figure 5: Decision tree with class names.

An object is classified by beginning at the root of the tree, finding the value of the tested attribute in the given object, taking that branch and continuing in the same fashion until a leaf is reached. Notice that classifying a particular object may only require testing a small number of its attributes. In the example above, it is not necessary to determine the value of the "height" attribute. The strength of the ID3 algorithm is the ability to identify and discard irrelevant attributes for problem solving. The ID3 algorithm will always work provided there is no "noise" in the data. Noise will be present if some of the examples had missing values, or if examples with identical attributes belonged to different classes.

A somewhat different approach to problem solving is adopted by AQ15 which uses a version of first order predicate calculus that has been modified to express inductive generalizations more easily. AQ15 describes observations in terms of selectors. A selector consists of a term followed by a relational symbol (<, >, <=, >=, <>, =) followed by a value. For example a selector may be "size<=15" or "color=blue". A combination of the description of an observation and the classification of the observation as either a positive or negative example of the target concept is called an example. AQ15 then uses one or more examples to form a hypothesis and then begins an iterative process, called STAR, which formulates and modifies hypotheses until all

positive examples of a concept are included, and all negative examples are excluded. The results are then displayed to the user.

One of the most important features of AQ15 is the types of domains that it allows for data. Nominal, numerical, and structured nominal values may all be used. An example of a nominal domain is the set {blue, red, green} for the attribute "color". Numeric data can be either integers or intervals of integers. A structured nominal domain has extra values in addition to the feature values present in the examples. These extra values (which are nominal) are values to which a system may generalize. The set of all (including the extra) values of a structured nominal attribute may be ordered by their degree of generality in a generalization tree, such as the one shown in Figure 6.

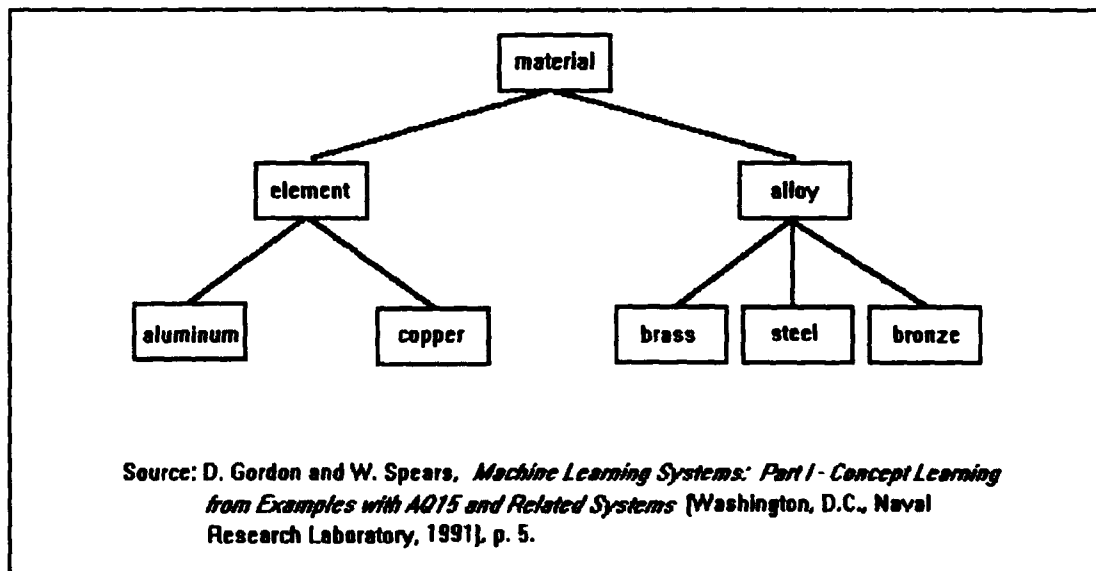


Figure 6: Generalization tree of descriptive terms.

The root of the generalization tree (i.e. the value at the top of the tree) implies every value of the feature. Roots are labeled with the name of the feature. The leaves of a generalization tree (i.e. the values at the bottom of the tree) are the attribute values which are present in the examples. For example, using the figure below, if all members of the positive examples had the value "brass", "steel", or "bronze" for the attribute "material" then the system could generalize that for positive examples, "material = alloy." The ability to generalize attributes is an important feature of AQ15.

IXL draws on the underlying theory of ID3 and AQ15 in the following manner:

- If the problem can be easily classified, then an ID3-like tree is produced
- AQ15-like methods of structure representation are supported and the program can generalize to attribute values which are not included in the data set.

Using these concepts as a background, we will examine more specifically how IXL learns concepts.

IXL is composed of five modules. The actual discovery process occurs in two of them, the Discovery Module and the Induction Engine. The Discovery Module searches the database for relationships and patterns. The search may be guided by user defined criteria (supervised) or allowed to search for any interesting patterns (unsupervised). The user may also filter unwanted information by specifying discovery

parameters, maximum rule length, and level of interest in the individual attributes. IXL's use of statistical methods occurs primarily at this stage. Several conventional statistical techniques are used to identify relationships, correlations and structure. These techniques are combined with logical analysis in order to guide search and interpret results automatically. The non-statistical portions of IXL are primarily rule based and written in the proprietary Intelligence/CompilerTM system.

The correlations discovered by the Discovery Module are then used by the Inference Engine to generate rules of knowledge, expressed in terms of user defined concepts and criteria. Correlations are identified and rules generated using both conventional statistical and non-statistical methods. [Ref.12, p.6]

2. Output Format

IXL summarizes results as logical rules. Logical rules are simple to read and do not require interpretation by an person knowledgeable in statistics. Figure 7 shows a rule which was generated from a database compiled by a computer disk drive manufacturer.

```

% Rule 15
CF = 77
"OPERATOR" = "12345"
IF
  "Error_Code" = "70"
;
% Margin of Error: 6.2%
% Applicable Percentage of Sample: 5.4%
% Applicable number of records: 180

Source: IXL User's Manual
        (Los Angeles, Ca., IntelligenceWare, Inc.,
        1990), p.1-15

```

Figure 7: Example IXL rule.

This rule states that about 77% of the manufacturing of the faulty drives of Error Code 70 is monitored by Operator number 12345. Therefore it is possible that Operator number 12345 is responsible for the faulty disk drives of Error Code 70.

The Margin of Error of 6.4% means that the Confidence Factor (CF) may be in error by as much as 6.4%. The actual CF is therefore between 70.6% and 83.4% (77% plus or minus 6.4%). The maximum margin of error used by IXL may be specified by the user. The confidence factor is analogous to type II error. If the margin of error increases, then the allowance for type II error increases. If the allowance for type II error increases, then each individual rule needs to meet less stringent statistical standards before meeting the minimum to be reported, i.e. a lower critical value. Rules

with lower critical values tend to be applicable to smaller portions of the database. In general, a larger margin of error will produce more rules which are relevant to only a small portion of the database.

The Applicable percentage of sample refers to the percentage of the database records that the "IF" condition satisfies. In this case, there are about 5.4% (180 records) of the database for which "ERROR_CODE = 70" is true.

Rules may be composed of multiple concepts. For example, Figure 8 show how a rule generated from a database of baseball statistics might read:

```
CF = 65
  "Home Runs"
IF
  ".293" <= "BATTING AVERAGE" <= ".342"
AND
  "128" <= "RBI" <= "176"

Source: IXL User's Manual
        (Los Angeles, Ca., IntelligenceWare, Inc.,
        1990), p. 1-19
```

Figure 8: Sample IXL rule with two identifying concepts.

Up to 7 different conditions may be included in the "IF-AND" statement. The maximum rule length is a user specified parameter.

Once the user is satisfied with the rules produced by IXL, the rules can either be used directly or as an input to an expert system using the Intelligence/Compiler system.

III. DATA AND METHODOLOGY

A. INTRODUCTION

The primary purpose of this thesis is to evaluate the effectiveness of the two different machine learning programs. To accomplish this, the two programs examined must be compared not only to each other, but also to some known standard. Currently the most widely accepted, and popular, method of analyzing personnel data sets is multivariate regression. Therefore the primary basis of comparison is how the results generated using the machine learning programs compare with those generated by using conventional regression analysis techniques. The general methodology for this thesis is:

1. Acquire a large manpower data set.
2. Randomly divide the data into a training data set and a test data set.
3. Obtain the regression equation which was developed using the training data set.
4. Use the training data set to develop an AIM network and generate IXL rules.
5. Use the test data set to evaluate the predictive capabilities of the regression equation, AIM, and IXL.
6. Compare the results of all three programs and evaluate any other strengths or weaknesses which become evident during the model development and evaluation process.

The rest of this chapter examines the data set and the methodology used to develop the different models.

B. DATA

1. Data Source

The data used were primarily extracted from the 1985 DOD Survey of Officer and Enlisted Personnel. The survey data were matched with personnel records, using social security numbers, to obtain information on respondents' active duty status in 1989.

The 1985 survey was conducted by the Defense Manpower Data Center in response to a request from the Deputy Secretary of Defense for Force Management and Personnel. The primary purpose of the survey was to provide information which could be used by the armed services to improve retention and readiness. Table 1 describes the nine sections of the survey. The survey was fielded to a sample of 132,000 active duty officers and enlisted personnel worldwide from all of the United States military services. Personnel with less than 4 months of active duty service were excluded.

**Table I 1985 DOD SURVEY OF OFFICER AND ENLISTED PERSONNEL:
TOPIC AREAS**

<u>Section</u>	<u>Questionnaire Topic Area</u>
1	<u>Military Information</u> --Service, Paygrade, military occupation, term of enlistment
2	<u>Present and Past Locations</u> --length of stay, expected stay, and problems encountered at present and past duty stations
3	<u>Reenlistment/Career Intent</u> --expected years of service, expected rank when leaving the service, and probable reenlistment behavior
4	<u>Individual and Family Characteristics</u> --basic demographics such as age, sex, and marital status
5	<u>Dependents</u> --basic demographics from Section 4, and whether or not dependents were handicapped
6	<u>Military Compensation, Benefits, and Programs</u> --benefits received for military service, and availability and satisfaction with family programs
7	<u>Civilian Labor Force Experience</u> --members' civilian work experience and previous earnings
8	<u>Family Resources</u> --household's civilian work experience and earnings, and non-wage or salary sources of earnings
9	<u>Military Life</u> --satisfaction with various aspects of military life, including pay and allowances, interpersonal environment, and benefits

Source: 1985 DoD Survey of Officers and Enlisted Personnel

2. Thesis Data Set

The specific data set used in this thesis was obtained from Dr. George Thomas and Kathryn Kocher of the Naval Postgraduate School. The data consist of selected variables

and observations from the 1985 DOD survey. The sample was limited to Navy, male, enlisted personnel with 24 to 72 months of active duty service, and in pay grades E3 - E6. Respondents included in the sample also had to be within 3 years of their end of active obligated service at the time of the survey. The 3 year limitation was imposed in order to ensure that each member of the sample had made at least one reenlistment decision prior to 1989.

Personnel who were older than 30 years of age at the time of the initial enlistment were also excluded. The 30 year age cutoff was imposed because there is evidence that personnel in this age group are making a final lifetime career decision upon initial enlistment, and therefore their reenlistment behavior is significantly different from the general Navy enlisted population. Finally, any observations which contained missing or undefined variables were omitted.

After defining the population for the data set, specific variables were selected which were known to be relevant to a decision to reenlist or not reenlist. The final data set had 780 observations and 17 variables. This data set was randomly split into a training data set with 680 observations and a test data set with 100 observations.

3. Variable Definitions

a. Dependent Variable (STATUS)

The dependent variable STATUS measured the reenlistment behavior of the sample member. Because every member had made a reenlistment decision prior to 1989, a member still on active duty must have reenlisted. STATUS equals one if the member was still on active duty in 1989, otherwise the variable was equal to zero.

b. Independent Variables

There are 16 independent variables included in the model. These variables fall into four general categories: Demographics, Military Characteristics, Educational Level, and Satisfaction with Military Lifestyle and Benefits. Each variable is described in the sections below. A hypothesis for the effect of each variable on STATUS is also given.

(1) Demographic Variables

a) Age Upon Entering Active Duty Status. ENTRYAGE is the member's age at the time of initial enlistment. ENTRYAGE was computed by subtracting the number of months on active duty from reported age at the time of the survey. As a member's age at enlistment increases, the probability that he had worked in the civilian labor market increases. The decision to enlist indicates that he disliked his civilian job in comparison to his perceived opportunity in

the Navy. An individual who is older at the time of enlistment also has fewer years to establish a second career after completing an initial enlistment. Therefore, ENTRYAGE is expected to have a positive effect on STATUS.

b) Ethnic Background. The effects of ethnic background are measured using three dummy variables WHITE/OTH, BLACK, and HISPANIC. A dummy is coded as a one if the member is from the appropriate ethnic group. The HISPANIC variable includes only non-black hispanics (black hispanics are included in BLACK), and a person could only be a member of one group. Past studies have shown that ethnic minorities reenlist at a higher rate, because of a perceived lack of opportunity in the civilian labor market. Therefore, BLACK and HISPANIC are hypothesized to have a positive effect on STATUS, as compared to WHITE/OTH.

c) Family and Marital Status. The effects of family and marital status are measured using four dummy variables: Single No Children (SNC), Single With Children (SWC), Married No Children (MNC), and Married With Children (MWC). The variable which described the member's status was coded as one, the other three variables were coded zero. As the number of dependents a member is responsible for increases, so does the aversion to risk. Because there is rarely a guarantee of a job in the civilian market when a member leaves the military, leaving the Navy is more risky than reenlisting with a guaranteed paycheck. Therefore,

SWC, MNC, and MWC are hypothesized to have a positive effect on STATUS, as compared to SNC.

(2) *Military Characteristics*

a) Rank. Three dummy variables; E3, E4, and E5/6, were used to measure the effects of rank on reenlistment. The ranks E5 and E6 were combined because these members are usually in their second or subsequent enlistment and exhibit similar retention characteristics. Increased rank leads to increased pay, benefits, and responsibilities, decreasing the incentive to leave the Navy. Therefore, increased rank is hypothesized to have a positive effect on STATUS.

b) Military Occupation. The effects of different occupations are measured using a dummy variable which indicates if the member is in a technical occupation (TECHOCC). If the member was in an occupation in one of the following general categories he was considered to have been in a technical occupation and the TECHOCC variable was coded one, otherwise the variable was coded zero:

1. Electronic Equipment Repair
2. Communications and Intelligence
3. Medical and Dental
4. Other Technical Fields

Members in technical occupations have skills which are valuable in the civilian labor force. Because they have greater opportunities outside of the military than members without technical skill they were expected to leave the Navy

at a higher rate. Therefore, TECHOCC is hypothesized to have a negative effect on STATUS.

c) Probability of Finding a Good Civilian Job.

The effect of whether the member believed he had a good opportunity for a civilian job was measured by the variable CIVJOB. CIVJOB was coded one if the member believed he had a good opportunity and coded zero if he did not believe he had a good opportunity. If a person believed that he had a good opportunity for a civilian job, then he would probably be less likely to stay in the military. Therefore, CIVJOB is hypothesized to have a negative effect on STATUS.

(3) *Educational Accomplishment.* The effect of educational accomplishment is measured using a dummy variable, High School Certificate Holder (HSCERT). If a member had a high school diploma HSCERT was coded as zero. If the member had a GED certificate, a high school completion/attendance certificate, or a home study diploma, then HSCERT was coded as one. If the member did not have a high school diploma or high school certificate equivalent, then he was dropped from the sample. Members without a high school diploma should be at a significant disadvantage in the civilian labor market. Therefore, HSCERT is hypothesized to have a positive effect on STATUS.

(4) *Satisfaction with Military Lifestyle and Benefits.* A significant portion of the 1985 DOD Survey of Officer and Enlisted Personnel is concerned with a member's satisfaction with military lifestyle and benefits. Satisfaction with the military should have a significant impact on a member's decision to reenlist. However, there are some significant problems with using the satisfaction variables directly. The most important problem is that the variables are highly correlated with each other. Multicollinearity among the independent variables does not change the overall predictive capability of a model, but it does affect the significance levels of the explanatory variables and the computed partial effects. Because explanation is often as important as prediction, multicollinearity reduces the overall effectiveness of a model. One solution to the problem of multicollinearity between the independent variables is factor analysis. Factor analysis will yield new explanatory variables which are uncorrelated with each other. Factor analysis was undertaken using thirteen satisfaction variables from the 1985 DOD Survey. Two underlying dimensions were identified. Table 2 shows the rotated factor pattern scores for the thirteen variables.

Table II ROTATED FACTOR PATTERN SCORES

<u>Satisfaction Variables</u>	<u>FACTOR1</u>	<u>FACTOR2</u>
Overall Job	0.71266	.
Work Conditions	0.62253	.
Job Training	0.55176	.
Job Stability	0.54597	.
Co-Workers	0.51141	.
Job Security	0.49178	..
Promotions	0.47001	.
Personal Freedom	0.46376	.
Ability to Serve Country	0.42604	.
Family Environment	0.41481	0.37981
Friendships	0.36824	.
Moving	0.35458	.
Medical Care	.	0.76467
Dental Care	.	0.69765
Commissary Services	.	0.50460
Retirement Benefits	.	0.43947
Pay	0.38413	0.43609
VEAP Benefits	.	0.41571

Note: Values less than 0.3 have been printed as '.'

Factor1 is heavily influenced by satisfaction with military work and lifestyle variables. Therefore this factor was renamed MILLIFE. The second variable, FACTOR2 was primarily influence by military pay and benefits. This

variable was renamed MILBENE. If satisfaction with any facet of military lifestyle or benefits increases then a member would have more incentive to stay in the military. Therefore both MILLIFE and MILBENE are hypothesized to have a positive effect on STATUS.

C. METHODOLOGY

1. Multivariate Logistic Regression Analysis

Multivariate regression was used to estimate the relationship between the dependent variable, STATUS, and the independent variables identified in the previous section. This portion of the analysis used the 680 observation training data set. The specific estimation technique was binomial logistic regression. This technique is the most suitable for estimating a dichotomous dependent variable, such as STATUS. Logistic regression provides the following relationship:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_n X_{ni} + \epsilon$$

where P_i is the probability that the i th person will make a given choice, in this case to reenlist and remain on active duty, given his set of explanatory variables (X_1, X_2, \dots, X_{17}). The dependent variable in this equation is the logarithm of the odds that a particular choice will be made. The appeal of the logit model is that it transforms the problem of predicting the probabilities within the (0,1)

interval to the problem of predicting the odds of an event's occurring within the range of the real line.

After the logistic regression equation was estimated the coefficients were used to evaluate each of the 100 test cases. The probability that an individual would reenlist was estimated using the equation:

$$P_i = \frac{1}{1 + e^{-\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_n X_{ni}}}$$

where the β 's were the estimated coefficients and the X's were the actual values of the independent variables for the test individual. Using a probability cutoff of .5, the number of correct predictions was computed.

2. AIM Model

AIM has two primary network synthesis parameters: complexity penalty multiplier (CPM) and number of layers. The default settings are CPM=1 and number of layers = 4. In order to evaluate the effect that each parameter has on the network development, 5 different CPMs (.5, .8, 1.0, 1.2, and 1.5) and 3 different number of layers (3, 4, and 5) were utilized. Using these values, 15 different networks were synthesized.

After the networks were developed, the test cases were evaluated. AIM has the capability to do this directly. Using a probability cutoff of .5, the number of correct predictions was computed.

3. IXL Rules

The data set for IXL was the same as the one used for regression analysis and AIM; however, the format was slightly different. The IXL manual recommends that data be converted into variables with descriptive, qualitative values, if possible [Ref.13, p.1-19]. Because IXL is capable of utilizing descriptive, categorical data directly without dummy variables, several of the variables were combined to create a new data set. The new data set has the same information, but in a format that is easier for IXL to evaluate. If the data were not transformed, IXL would have had more difficulty discovering rules and the discovery process would have taken longer, although the final rules would have been basically the same.

The dependent variable STATUS was recoded with values "Still in Military" and "Not in Military". The ethnic variables WHITE/OTH, BLACK, and HISPANIC were combined to create the variable ETHNIC. The ETHNIC variable was given a descriptive coding which was either "White", "Black", or "Hispanic". The four marital and dependent status variables were combined to create the descriptive MARITAL variable, with values: "Single No Children", "Single With Children", "Married No Children", and "Married With Children." TECHOCC was recoded to either "In Technical Occupation" or "Not in Technical Occupation".

The HSCERT variable was used to create the variable HSSTATUS with values "High School Diploma Grad" or "High School Certificate". The MILLIFE and MILBENE variables were converted from numeric variables with continuous values between negative one and one to a 5 stage descriptive variable with ranges from "Dissatisfied" to "Satisfied". The exact values are given in Table 3.

Table III VALUES FOR NEW MILLIFE AND MILBENE VARIABLE

<u>OLD VALUE</u>	<u>NEW VALUE</u>
X <= .5	"DISSATISFIED"
-.5 < X < -.2	"SOMEWHAT DISSATISFIED"
-.2 < X < .2	"NEUTRAL"
.2 < X < .5	"SOMEWHAT SATISFIED"
X > .5	"SATISFIED"

IXL has several parameters which must be set prior to rule discovery. There are very few "default" settings. The parameters used were:

- Sampling Percentage = 100%
- Maximum Rule Length = 7
- Minimum Rule Confidence = 85%
- Maximum Margin of Error = 10%
- Minimum Group Size = 20%
- Level of Significance = 60
- Minimum Generality = 5%
- Maximum Generality = 100%
- Generality Increments = 100%
- Maximum Run Time = 9999 minutes

According to the IXL Users Manual, these discovery parameters would yield a reasonable number of significant rules to be evaluated.

After IXL generated its rules, each rule was separately evaluated using the test data set. The rules in combination were also evaluated to determine if some combination of rules was effective for predicting reenlistment behavior.

After determining which rules were most effective for prediction, the test data set was used to determine the number of correct cases that IXL predicted.

D. EXPANDED VARIABLE DATA SET

When synthesizing a network, AIM discards variables which do not contribute significantly to the solution of the models. If AIM is able to do this effectively, then a user would be able to synthesize an AIM network from a large data set, with many potential independent variables, without determining the theoretical relationship between the dependent variable and each of the independent variables. In this case AIM would be used to identify variables which have an impact on the dependent variable and, after the network is synthesized, the identified variables could be evaluated for theoretical validity. Because determining which variables should be

included in a model is time consuming and labor intensive, a-priori identification of independent variables that affect the dependent variable could save valuable time and scarce resources.

To examine AIM's capability to identify relationships within a larger data base, 10 additional variables were added to the data base. These variables were added without determining potential effects on an individual's decision to reenlist. A network was synthesized using both the 17 variables included in the previous data set and the new variables. The ten new variables were:

- SPACTIVE - This variable was coded one if the individual's spouse was also on active duty in any of the Armed Forces, otherwise it was coded zero.
- SEATIME - The value for this variable was the number of months that a member has spent at sea during his career.
- OSEATIME - The value for this variable was the number of months that an individual spent at an overseas duty station.
- INCOME - INCOME was a continuous variable with the value equal to total family income.
- PCSMOVE - This variable had a value equal to the number of permanent change of station (PCS) moves that an individual had made in his career.
- MOMSED - MOMSED is equal to the number of years of education that an individual's mother had completed.
- OFDTYJOB - OFDTYJOB is the number of hours per week that an individual spends at a civilian job during his off duty hours.
- CIVJOBBOF - CIVJOBBOF was coded one if the individual had received a civilian job offer in the previous year, otherwise CIVJOBBOF was coded zero.

- MILHOUR - MILHOUR was the time of day, measured using a 24 hour clock, that the individual began completing the 1985 DOD survey.
- DEBT - The value of the DEBT variable was dependent upon the total amount of outstanding debt, excluding mortgages, that an individual had. The variable was coded between one and seven. All seven codes were used in the AIM model. The codings, and amount of outstanding debt, were:

<u>CODE</u>	<u>DEBT</u>
1	NONE
2	\$1 - \$499
3	\$500 - \$1999
4	\$2000 - \$4999
5	\$5000 - \$9999
6	\$10000 - \$14999
7	\$15000+

IV. RESULTS

A. LOGISTIC REGRESSION

A binomial logistic regression equation was estimated to use as a base against which each of the other two software programs could be compared. The regression equation was estimated using the 680 observation training data set. The equation was then evaluated using the 100 observation test data set. The specific equation used for evaluation was:

$$P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \dots + \beta_n X_{ni})}}$$

where P_i is the probability that the i th individual reenlisted in the Navy, the β s are the estimated coefficients, and the X_i s are the values for the independent variables for the i th individual. If $P_i \geq .5$ then the equation predicts the individual will reenlist, otherwise it predicts the individual will leave the Navy. The prediction was compared to the STATUS variable to determine if the equation had correctly predicted reenlistment behavior. The coefficients, significance levels, and equation goodness of fit statistics are shown in Appendix A. The regression equation correctly predicted 72 of the 100 test cases.

B. AIM

Fifteen different AIM networks were synthesized using the 680 observation data set. The 100 case data set was used to evaluate the predictive capabilities of each network. The evaluation procedure was similar to that for the regression analysis. AIM computed a probability that the individual would reenlist and, using a cutoff value of .5, the number of correct predictions was computed. This procedure was used for each of the 15 networks. The graphical representations of the default parameter network (CPM = 1.0 and number of levels = 4) and the network with the best predictive capability (CPM = .5 and number of levels = 4) are shown in Appendix B. The numbers of correct predictions for the AIM networks are shown in Table 4.

Table IV NUMBER OF CORRECT PREDICTIONS FOR AIM NETWORKS

		COMPLEXITY PENALTY MULTIPLIER				
		<u>0.5</u>	<u>0.8</u>	<u>1.0</u>	<u>1.2</u>	<u>1.5</u>
NUMBER	3	71	71	71	71	71
OF	4	72	71	69	71	69
LAYERS	5	71	69	69	70	69

Unlike regression analysis, AIM does not generate an overall "goodness-of-fit" statistic for its networks, so no statement

can be made about the statistical significance of the networks, or of the individual variables.

C. IXL

IXL "discovered" 34 rules in the training data set. The rules and the number of correct predictions for each rule are presented in Appendix C. Individually Rules 13, 19 and 23 were the best predictors, each correctly predicted 45 individuals. Rules 17, 24, and 27 were the worst predictors, correctly predicting 35 individuals. To evaluate the collective predictive capability of the rules, a correct prediction for IXL was defined as: 17 of the 34 rules correctly predict an individual's reenlistment behavior. Using this criterion, IXL correctly predicted 36 of the 100 test cases.

To determine if there was a particular subset of rules that was significantly more accurate for predicting reenlistment a new 680 observation data set was created. This data set included the dependent variable STATUS and 34 independent variables. Each of these independent variables corresponded to one of the IXL rules. The variable was coded zero if the rule predicted that the individual left the Navy, and was coded one if the rule predicted that the individual reenlisted. A binomial logistic regression equation was then estimated using the new data set. Rules which were significant at the .10 level or better were retained and the

others discarded. Six rules (Rules 7, 12, 13, 15, 17, and 28) were significant at the .10 level, and these were used to evaluate the test data set. A "correct" prediction was redefined as 3 of the 6 remaining rules correctly predicting reenlistment behavior. Using this criterion, 42 of the 100 test cases were correctly predicted.

Because IXL does not discover a "model" there is no overall goodness-of-fit statistic. IXL does give a confidence factor and a margin of error for each of the discovered rules. This information is included in Appendix C.

D. EXPANDED VARIABLE AIM MODEL

The 27 variable expanded data set was used to create an AIM network. The default discovery parameters (CPM = 1.0 and number of levels = 4) were used. The graphical representation of the expanded AIM network is shown in Appendix D. Using the same criteria as the other AIM networks, the expanded network correctly predicted 72 of the 100 test cases.

V. DISCUSSION

A. INTRODUCTION

Often models are developed to guide decision makers. Rarely are the costs of making an incorrect decision the same for all alternatives. If an individual's effectiveness is highly dependent upon expensive training or experience, for example, it may be extremely expensive to allow that person to leave the Navy. In this case, it would be much more expensive to the Navy if the model incorrectly predicts that a person is going to reenlist and have him leave the service, than it would be to incorrectly predict that the person would leave, pay him a bonus and find out that he would have stayed in any case.

It may also be important to know how small changes in an independent variable influence a person's decision to reenlist. For example, it may be important to know that a \$1000 reenlistment bonus will have a greater influence on an E3 than on an E5.

Because the costs of making incorrect decisions differ, this chapter examines more completely the results obtained from the binomial logistic (logit) regression equation and the AIM networks to determine if the two software programs are

significantly different for predicting behavior. Comparisons between the programs will be made in three areas:

- Predictions for individual observations,
- The effects of small changes in the independent variables,
- Other strengths and weaknesses.

IXL is not included in the comparisons for the first two areas for two reasons: the format of the output is significantly different from the other two programs, making comparisons difficult; and, based upon the results presented in the previous chapter, IXL does not predict reenlistments as well as the other two programs.

B. PREDICTION

1. Overall Comparison

Sixty-two of the 100 individuals in the test set were predicted correctly both by the logit equation and by all 15 AIM networks. An additional 8 observations were correctly predicted by the logit equation and by some (between six and 13) of the AIM networks. Only two observations were correctly predicted by logit and by none of the AIM networks; and only one observation was correctly predicted by all of the AIM networks and not by the logit model. For three of the observations the logit model was incorrect and some of the AIM models were correct. For 23 observations the logit equation and all 15 AIM networks were incorrect.

2. Best AIM Network

A direct comparison can also be made between the logit model and the best AIM model, both of which correctly predicted 72 individuals. Sixty-nine of the correct predictions, and 25 of the incorrect predictions, were the same for the two programs. Each of the programs correctly predicted three of the remaining six individuals.

Both the logit equation and AIM had difficulties predicting which individuals were going to reenlist. Thirty-three of the individuals in the test data set reenlisted. Of those 33, logit correctly predicted 13 and the best AIM network predicted 14. Seventeen of the 33 were not correctly predicted by any of the AIM networks.

In summary, there does not appear to be a significant difference in the predictive capabilities of the logit equation and AIM. Both predict approximately the same number of people, and primarily the same individuals.

C. EFFECT OF CHANGES IN INDEPENDENT VARIABLES

1. Introduction

Another important capability of regression equations is the ability to evaluate the effect that a small change in an independent variable will have upon the dependent variable. This is important for evaluating the possible effects of a change in policy. If a decision maker changes a policy that affects one of the independent variables it is important to

know not only that the change will have an effect, but the magnitude of the effect.

For a linear regression equation analysis is simple, a one unit change in the independent variable will create a change in the dependent variable equal to the coefficient for that independent variable. Because the equation is linear, a one unit change in an independent variable will always have the same impact on the dependent variable.

For a binomial logistic equation analysis of partial effects is more difficult. The coefficient represents the impact of a one unit change in the independent variable on the log of the odds of a given choice, in this case the decision to reenlist (STATUS=1), not on the probability itself.

To evaluate the effect of a one unit change in an independent variable it is necessary to define a "base case" individual. All effects created by changes in independent variables can then be evaluated using the base case as a comparison. Specifically, the logistic regression equation in Chapter IV is used to compute the probability of reenlisting for the base case. After determining the base probability, a one unit change in one independent variable is made, holding all other independent variables constant, and a new probability is computed. The difference in probability is the effect of the one unit change in that independent variable on the base case individual. Because the logit equation is not linear, the amount of change is dependent upon the

characteristics of the base case individual. A different base case will lead to a different effect for a one unit change in the independent variable. Also, a change of one additional unit for the same independent variable will not result in the same change in probability as the first one unit change.

Another important factor when evaluating the effect of a change in an independent variable is the statistical significance level of the variable. If the statistical probability level of the independent variable is not less than or equal to a preselected maximum value (usually .05), then the regression equation determined that the independent variable had no effect upon the dependent variable. Regardless of the coefficient and computed partial effect, if the independent variable is not significant the computed effect is equal to zero.

AIM also has the capability to evaluate the effects of changes in independent variables. Using the program's "Query" function, a user can enter values for each of the independent variables and AIM will compute the probability of reenlistment. This function allows the user to evaluate changes in probabilities in much the same manner as comparing the probabilities for a logistic equation.

2. Computed Effects

Because logistic regression is the preferred method of computing the effects of changes in the independent variables,

and because these results are often used to guide decision makers in creating policy, it is important to determine if AIM predicts similar effects to logistic regression.

To estimate the change in reenlistment behavior of the "base case" individual, it is necessary to first define a base case enlistee. For the purposes of this thesis a base case enlistee is defined as having the following characteristics: entry age equal 19, E3, white, single with no children, and all other variables equal zero. These values were chosen because they represent the "average" individual, i.e. the mean value for continuous variables, and the modal value for categorical variables.

The predicted changes computed by the logistic regression, the default AIM network, and the best AIM network (in terms of predictive capability) are shown in Table 5. These changes assume that all other variables are held constant.

Table V CHANGE IN PROBABILITY OF ENLISTMENT FROM BASE CASE.

VARIABLE	LOGISTIC MODEL	DEFAULT AIM	BEST AIM
ENTRYAGE=20	.01	.01	.01
E4 = 1	.13*	.12	.10
E56 = 1	.21*	.22	.25
BLACK = 1	.14*	.14	.03
HISPANIC = 1	-.02	-.11	-.17
SWC = 1	.05	.03	.09
MNC = 1	.17*	.09	.14
MWC = 1	.17*	.09	.15
TECHOCC = 1	.04	.02	.01
CIVJOB = 1	-.09*	-.03	-.01
HSCERT = 1	.04	.02	.01
MILLIFE = 1	.06*	0	.02
MILBENE = 1	.03	-.03	.02
BASE PROB	.22	.23	.20

* -significant at the .05 level

The MILLIFE and MILBENE variables are included in Table 4. Because these two variables are a composite of other variables created by factor analysis, what would cause a one unit change in a satisfaction variable is not immediately obvious. A one unit increase in factor score occurs when the values of the underlying variables increase sufficiently that when they are subjected to factor analysis the resulting factor score (the

MILLIFE and MILBENE variable) is one standard deviation greater than before the change. These variables cannot be manipulated directly by policy makers, but they do give an indication of how satisfaction with the military can affect the decision to reenlist.

3. Comparison

The results obtained from the logit equation and the best AIM network are very similar. All of the effects have the same sign and, except for the ethnic variables, the differences between the two are less than .05. The difference in the Hispanic variable is particularly large. The direction of the effect of HISPANIC on reenlistment is also opposite to the effect predicted in the methodology chapter. Because the Hispanic variable in the regression equation is not statistically significant, the computed effect is actually zero and the negative sign can be disregarded.

Both of the AIM networks compute large negative effects for HISPANIC. Mehay found that Hispanics were less likely to enlist initially in the military than whites or blacks [Ref.14, p.16]. This may indicate a cultural bias against military service. He also found that Hispanics receive a positive economic return in the civilian labor from military experience [ibid., pg 13]. Greater opportunities outside of the military would lead to a lower reenlistment rate. These two factors may account for the

effects predicted by the AIM networks. It is not possible to explain the differences between the logit equation and AIM without doing further research on Hispanics to determine if the original prediction was in error or if some factor not specified in the model is affecting the HISPANIC variable.

Another useful comparison is between the logit model and the AIM network which was synthesized using the default values (CPM=1.0 and number of layers=4). The default values would be used primarily by inexperienced researchers, or in the first stages of research to identify relationships in the data set which could be used to guide further research.

The effects computed by the default network are also very similar to those computed by logistic regression. For the ethnic variables, this network's computed effects are actually closer to the logit equation than the best AIM network. The computed effects for MNC and MWC are not as close to the logit equation as the best AIM network, but they do indicate the proper direction, and that these variables have a large impact on the reenlistment decision. This information would be useful in identifying areas that need to be investigated in more depth. For the rest of the independent variables, the effects computed by the default network are closer to those computed by the logit equation than the best AIM network.

In general, the default network may be as useful as the best network for evaluating the effects of changes in the independent variables.

Unlike predicting reenlistments where there is a known actual behavior that the prediction can be compared to, there is no known "correct answer" when evaluating the effect that a change in one of the independent variables has upon the decision to reenlist. Using the results obtained from the logistic equation is appropriate because it is the most widely accepted method for predicting the magnitudes of these effects. In general, the effects of changes in the independent variables that AIM predicts are similar to those predicted by logistic regression. The best AIM model is closest to the logistic model, but the default network also provides good estimates. Both AIM models provide information which would be useful to decision makers or that would help guide further research.

D. EXPANDED VARIABLE AIM NETWORK

The expanded variable AIM network was evaluated using the same criteria as the other models, predictive capability and analysis of effects of changes in the independent variables.

The expanded variable AIM network, using default parameters, correctly predicted 72 of 100 test cases. This is the same number of cases as both the logistic regression equation and the best AIM model when the 17 variable data set was used.

The effects of small changes in the independent variables in the expanded variable model are very similar to those in the original models. Table 6 shows the results from the expanded AIM network as compared to both the default AIM network and the best AIM network from the previous section. The base case reenlistee for the new analysis has the following characteristics: entryage equal 19, E3, white, single no children, 27 months sea time, 9 months overseas time, family income equal \$14000, MILHOUR equal 1200, and DEBT category equal two. All other variables were equal to zero. The base case characteristics are taken from the training data set and are mean values for continuous variables and modal values for categorical variables.

Table VI CHANGE IN PROBABILITY OF ENLISTMENT FROM BASE CASE: EXPANDED VARIABLE NETWORK.

VARIABLE	EXPANDED AIM	DEFAULT AIM	BEST AIM
ENTRYAGE = 20	.01	.01	.01
E4 = 1	.16	.12	.10
E56 = 1	.23	.22	.25
BLACK = 1	.04	.14	.03
HISPANIC = 1	-.17	-.11	-.17
SWC = 1	.03	.03	.09
TECHOCC = 1	.01	.02	.01
CIVJOB = 1	-.02	-.03	-.01
HSCERT = 1	.01	.02	.01
MILLIFE = 1	-.06	0	.02
MILBENE = 1	.01	-.03	.02
SEATIME = 28	0	N.C.	N.C.
OSEATIME = 10	.01	N.C.	N.C.
INCOME = 15000	0	N.C.	N.C.
CIVJOB OF = 1	.05	N.C.	N.C.
MILHOUR = 1300	0	N.C.	N.C.
DEBT = 3	.02	N.C.	N.C.
BASE PROB	.21	.23	.20

N.C. = NOT COMPUTED IN THIS NETWORK

AIM determined that the four new independent variables which are not included in this table (SPACTIVE, PCSMOVE, OFDTYJOB, and MOMSED) did not contribute significantly to predicting STATUS, and did not include them in the network.

The effects computed by the expanded variable AIM network are very similar to those computed by the two original AIM networks, increasing the number of independent variables did not appear to have much impact upon the predicted effects for the variables already calculated. AIM also eliminated four of the new independent variables which were included in the model. Three of the remaining seven new variables were calculated to have no effect upon the base case individual's probability of reenlisting. The one variable included in the data set which would appear to have no theoretical validity, MILHOUR, was included in the model. However, changing MILHOUR from 1200 to 1300 yielded no change in the predicted reenlistment probability for the base case individual. Changing MILHOUR to 2300 increased probability of reenlistment by .01. The remaining three variables were calculated to have relatively small effects (less than .05).

Based upon the calculated results, AIM appears to have two very useful capabilities. The first is that number of independent variables included in the data set does not significantly influence the calculated partial effects of those variables actually included in the network. This allows a researcher to use a data set with many variables without having to be concerned about theoretical relationships o

about how each of the variables may influence the calculated partial effects of the other variables, and therefore the usability for policy guidance. The second capability is that AIM appears to eliminate many of the independent variables which do not have a significant effect upon the dependent variable. This capability allows the researcher to input many variables and let AIM determine which variables are significant. This would allow the researcher to focus attention on only those variables which are known to have an effect on the dependent variable, saving time and resources.

E. OTHER STRENGTHS AND WEAKNESSES

This section will examine some of the strengths and weaknesses of the two machine learning programs that were identified while conducting this thesis.

1. Documentation

a. AIM

The documentation provided with the AIM program was organized, clear, and well illustrated. Four sample data bases are provided. Three of these data bases have well documented tutorials in the AIM user's manual. The tutorials are of increasing complexity to assist a user in learning the different capabilities of the AIM program. No further training, references, or assistance from AbTech was required to develop the networks used in this thesis.

b. IXL

The documentation provided with the IXL does not meet the same standard as that provided by AIM. A tutorial, with example data sets, is provided. However, the data sets are relatively simple and do not explore the full range of the documented capabilities. The user's manual is subdivided into the same sub-modules as the program, with little reference to how the sub-modules interact. For example, one section of the manual describes the procedure to create new variables by algebraically manipulating the original data set variables. Another section describes how to remove certain variables from the data set to create a new data set. However, the documentation does not inform the user that if new variables are created, it is not possible to remove some of the original variables to create a new data set.

Another weak area is explanation of the user defined "discovery" parameters. IXL has many user-defined parameters which guide the program as it conducts the discovery process. The user's manual is particularly weak in explaining how each of the parameters affects the discovery process. The user must undertake a trial-and-error process to determine how each parameter affects the number and type of rules that are discovered.

In general, using any of the more sophisticated features of IXL requires the additional training or assistance from someone familiar with the program. The rules generated

for this thesis used the basic capabilities of IXL that were clear from the user's manual, without additional assistance.

2. Output Interpretation

a. AIM

Interpreting the output from the AIM network is relatively simple. AIM displays a graphical representation of the network. The user can easily determine which variables AIM used to synthesize the network and how they are related. Each of the nodes can also be examined to determine the equation used in that node.

In order to determine the predicted output for an individual; the user can use the "Query" function, enter the values for the independent variables, and AIM will return the predicted output. AIM also has an "Evaluate" function which will evaluate a user provided test data set with the same variables as the training data set and give a predicted output for each of the observations in the test set.

The Query and Evaluate functions are very easy to use and understand. This is in comparison to evaluating a logistic regression where the user must be familiar with regression equations and how to interpret the coefficients and statistical information, as well as knowing and understanding the probability equation so that coefficients and the values for the independent variables can be transformed into probabilities. AIM is superior to logistic regression for ease in interpreting results.

b. IXL

Ease of interpreting results is one of IXL's strengths. The output is presented as rules, which are easy to read and understand. The statistics included with the rules are relatively simple, and the user's manual gives adequate information so that a person unfamiliar with statistics can interpret the meaning of the statistics. IXL's rules are the easiest output of the three programs examined in this thesis to read and interpret.

3. Model and Variable Significance

One weakness of both AIM and IXL is the lack of statistics and tests to evaluate the significance of either the model or the individual variables. Regression results give statistics that allow the user to evaluate both the overall predictive capability of the model and to determine if individual variables significantly impact that capability. This evaluation allows the user to determine which variables are important in the model, and to focus his attention on those variables. It also allows a researcher to compare models and determine which of the models best predicts the dependent variable.

The AIM User's Manual states that if a variable does not contribute significantly to the predictive capability of the model it will not be included. However, it does not specify how the program determines significance, nor at what level it would eliminate a variable from the model. AIM does give some basic information for use to compare different

networks (e.g. average squared error), but it does not conduct statistical testing to determine if the overall network has predictive validity.

IXL provides statistical tests for each of its rules. It does not attempt to evaluate each variable. Because it does not produce a model, only a series of rules, it does not attempt to evaluate the overall statistical significance of its rules as an aggregate.

The lack of statistical tests for AIM and IXL make it difficult to determine, without further analysis, if the model has any predictive power. It also makes it difficult to focus on which variables are most important. When AIM eliminates a variable from the network, it has determined that the variable does not contribute to predicting the dependent variable; that the effect is zero. Determining that a variable does not contribute to predicting the dependent variable is equivalent to determining that the variable is not statistically significant and therefore the effect of that variable is equal to zero. Therefore, it may be safe to assume that a variable which is not included in the network does not have any statistical significance. However, it may not be safe to assume the opposite, that if a variable is included that it does have statistical significance at a level normally used by researchers. In general, the lack of statistical measures make the networks more difficult to interpret.

VI. CONCLUSIONS AND RECOMMENDATIONS

A. CONCLUSIONS

This thesis examined the predictive capabilities of two machine learning programs, AIM and IXL, and how these capabilities compared to the most commonly used standard, binomial logistic regression analysis. AIM had similar capabilities to logistic regression for predicting individual reenlistment behavior and for evaluating the effects of small changes in independent variables. IXL was not as accurate as AIM or logistic regression for predicting reenlistments and does not attempt to predict the changes in behavior that would result from small changes in the independent variables. AIM also appears to have the capability to identify the most important independent variables from a large data set.

AIM was found to be very easy to use, the documentation well written and illustrated, and the output easy to interpret. No special training was necessary in order to achieve the predictive capabilities illustrated in this thesis.

IXL was relatively more difficult to use than AIM, the documentation not as well written, but the output was easier to evaluate than that from the other two programs. While no special training was required to obtain the output that was used in this thesis, additional training and/or better documentation may allow a user to achieve superior results.

B. RECOMMENDATIONS

Based upon the results from this thesis, AIM should be further evaluated in an operational environment. This evaluation should compare AIM to currently utilized techniques, including regression analysis, to determine if the results found in this thesis are suitable to a wider range of applications.

Further research should be conducted to examine AIM capabilities and limitations. Similar comparisons to those done in this thesis should be conducted, using different data sets, to determine if the apparent effectiveness of AIM can be generalized to other personnel analysis areas. Another area that should be investigated further is AIM's ability to ignore independent variables which do not contribute significantly to its ability to predict the dependent variable. If the capabilities identified in this thesis are applicable to a wide variety of data sets and problems, then AIM may have enormous utility for researchers.

AIM and IXL are only two of the machine learning programs available. Other programs should be examined to determine what capabilities they have in the personnel research area.

APPENDIX A. LOGISTIC REGRESSION EQUATION

Table VII ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATORS.

VARIABLE	PARAMETER ESTIMATE	LEVEL OF SIGNIFICANCE
INTERCEPT	-2.1500	0.0055
CIVJOB	-0.6590	0.0021
ENTRYAGE	0.0450	0.2396
E4	0.6535	0.0107
E56	1.0031	0.0001
BLACK	0.6995	0.0077
HISPANIC	-0.0909	0.7754
SWC	0.2468	0.7004
MNC	0.8204	0.0003
MWC	0.8361	0.0001
TECHOCC	0.2409	0.1973
HSCERT	0.2402	0.2983
MILLIFE	0.3209	0.0017
MILBENE	0.1812	0.0851

-2 LOG L = 806.312, CHI-SQUARE = 83.709 (p=0.0001)

NUMBER OF CORRECT PREDICTIONS = 72

APPENDIX B. AIM NETWORKS

NORMALIZERS:

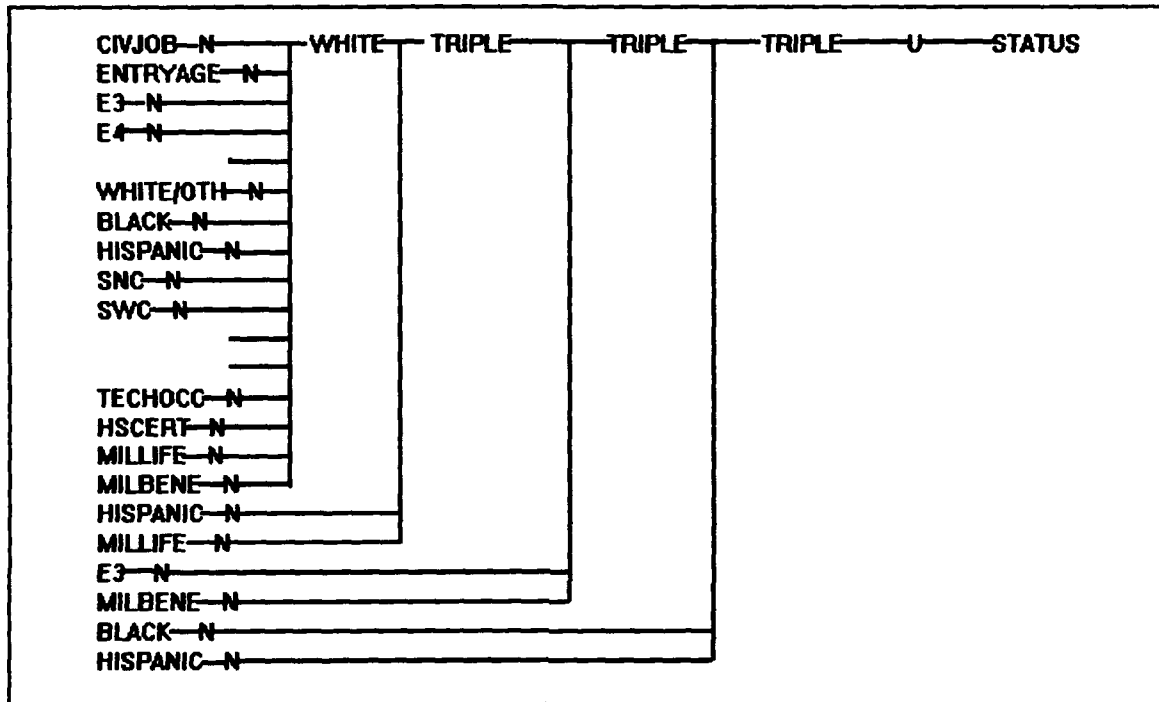


Figure 9: Default Network (CPM = 1 and number of levels = 4)

$$\begin{aligned}
 \text{CIVJOB } X_1 &= -2.08 + 2.56 * \text{CIVJOB} \\
 \text{ENTRYAGE } X_2 &= -8.64 + 0.448 * \text{ENTRYAGE} \\
 \text{E3 } X_3 &= -0.511 + 2.46 * \text{E3} \\
 \text{E4 } X_4 &= -0.794 + 2.05 * \text{E4} \\
 \text{WHITE/OTH } X_6 &= -1.97 + 2.48 * \text{WHITE/OTH} \\
 \text{BLACK } X_7 &= -0.373 + 3.05 * \text{BLACK} \\
 \text{HISPANIC } X_8 &= -0.299 + 3.63 * \text{HISPANIC} \\
 \text{SNC } X_9 &= -1.24 + 2.04 * \text{SNC} \\
 \text{SWC } X_{10} &= -0.134 + 7.59 * \text{SWC} \\
 \text{TECHOCC } X_{13} &= -0.673 + 2.16 * \text{TECHOCC} \\
 \text{HSCERT } X_{14} &= -0.439 + 2.71 * \text{HSCERT} \\
 \text{MILLIFE } X_{15} &= -0.011 + 1.13 * \text{MILLIFE} \\
 \text{MILBENE } X_{16} &= -0.0105 + 1.16 * \text{MILBENE}
 \end{aligned}$$

WHITE:

$$X_{17} = -0.116*X_1 + 0.0439*X_2 - 0.166*X_3 - 0.0779*X_4 - 0.239*X_6 \\ - 0.0949*X_7 - 0.173*X_8 - 0.184*X_9 - 0.0327*X_{10} + 0.0485*X_{13} + \\ 0.0402*X_{14} + 0.118*X_{15} + 0.0683*X_{16}$$

TRIPLES:

$$(1) X_{18} = 1.14*X_{17} + 0.397 * X_{17}^2 - 0.0665*X_{15}^2 + 0.211*X_{17}*X_8 \\ + 0.224*X_{17}*X_{15} + 0.104*X_{17}*X_8*X_{15} - 0.0609*X_{17}^3 \\ (2) X_{19} = 0.0994 + 0.935*X_{19} + 0.245*X_3 - 0.105*X_{16} - \\ 0.244*X_{19}*X_3 - 0.0701*X_3*X_{16} - 0.396*X_{19}*X_3*X_{16} - 0.0972*X_3^3 + \\ 0.0156*X_{16}^3 \\ (3) X_{20} = 1.01*X_{19} - 0.00971*X_8^2 - 0.093*X_{19}*X_7 + 0.00572*X_7^3$$

UNITIZER:

$$STATUS = 0.362 + 0.481*X_{20}$$

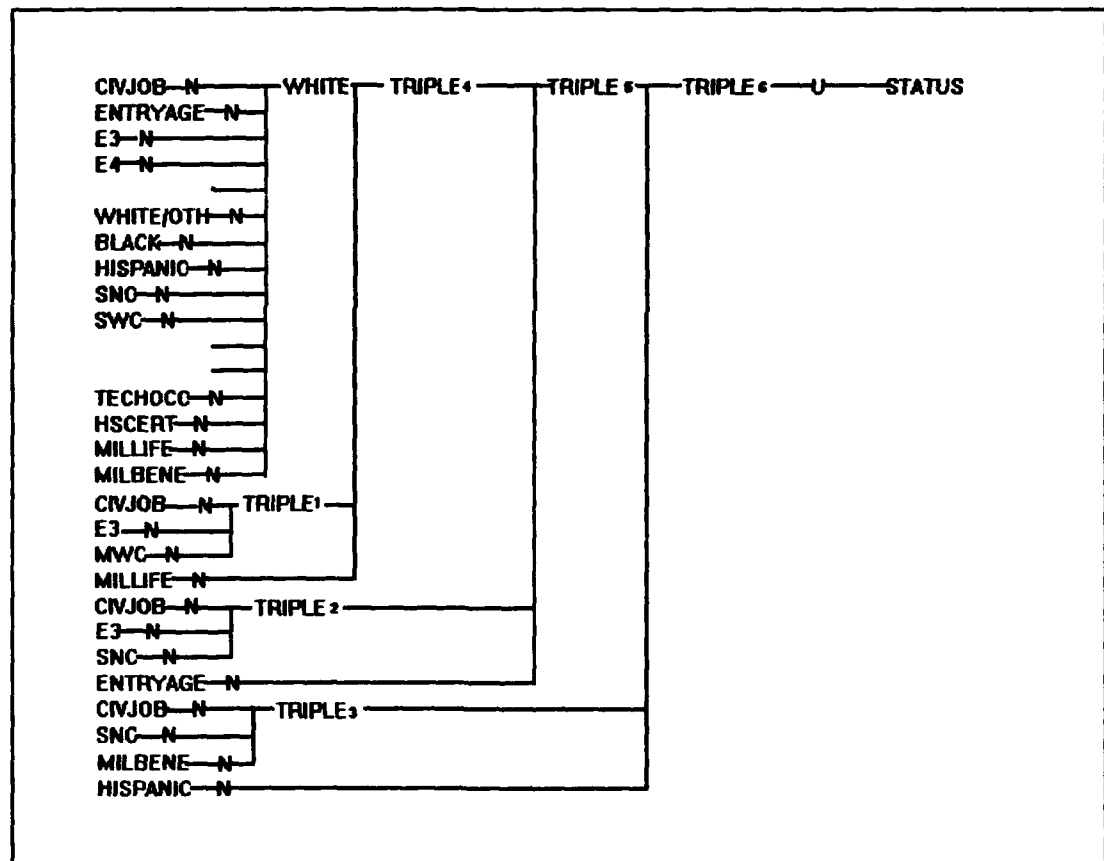


Figure 10: Best AIM Network (CPM = .5 and number of levels = 4)

NORMALIZERS:

$$CIVJOB X_1 = -2.08 + 2.56 * CIVJOB \\ ENTRYAGE X_2 = -8.64 + 0.448 * ENTRYAGE$$

$E3 X_3 = -0.511 + 2.46 * E3$
 $E4 X_4 = -0.794 + 2.05 * E4$
 $WHITE/OTH X_6 = -1.97 + 2.48 * WHITE/OTH$
 $BLACK X_7 = -0.373 + 3.05 * BLACK$
 $HISPANIC X_8 = -0.299 + 3.63 * HISPANIC$
 $SNC X_9 = -1.24 + 2.04 * SNC$
 $SWC X_{10} = -0.134 + 7.59 * SWC$
 $MWC X_{12} = -0.5 + 2.5 * MWC$
 $TECHOCC X_{13} = -0.673 + 2.16 * TECHOCC$
 $HSCERT X_{14} = -0.439 + 2.71 * HSCERT$
 $MILLIFE X_{15} = -0.011 + 1.13 * MILLIFE$
 $MILBENE X_{16} = -0.0105 + 1.16 * MILBENE$

WHITE:

$X_{17} = -0.116*X_1 + 0.0439*X_2 - 0.166*X_3 - 0.0779*X_4 - 0.239*X_6$
 $- 0.0949*X_7 - 0.173*X_8 - 0.184*X_9 - 0.0327*X_{10} + 0.0485*X_{13} +$
 $0.0402*X_{14} + 0.118*X_{15} + 0.0683*X_{16}$

TRIPLES:

(1) $X_{18} = 0.532 - 0.481*X_1 + 1.1*X_3 - 0.289*X_{12} - 0.215*X_{12}^2 +$
 $0.265*X_{12}^2 + 0.0784*X_1*X_3 + 0.0567*X_1*X_{12} - 0.106*X_3*X_{12} -$
 $0.0574*X_1*X_3*X_{12} - 0.41*X_3^3$
 (2) $X_{19} = 1.96 - 0.462*X_1 + 1.23*X_4 - 0.664*X_9 - 0.207*X_1^2 -$
 $1.11*X_9^2 + 0.08*X_1*X_4 - 0.047*X_1*X_9 + 0.0658*X_4*X_9 +$
 $0.0341*X_1*X_4*X_9 - 0.449*X_4^3$
 (3) $X_{20} = 1.11 - 0.194*X_1 - 1.49*X_9 + 0.0169*X_{16} - 0.0609*X_1^2 -$
 $- 0.691*X_9^2 - 0.0271*X_1*X_9 - 0.0504*X_1*X_{16} - 0.0142*X_9*X_{16} -$
 $0.108*X_1*X_9*X_{16} + 0.81*X_9^3 + 0.0152*X_{16}^3$
 (4) $X_{21} = -0.04 + 1.05*X_{17} - 0.147*X_{18} - 0.383*X_{17}^2 -$
 $0.282*X_{18}^2 - 0.0777*X_{15}^2 + 1.53*X_{17}*X_{18} + 0.541*X_{17}*X_{15} -$
 $0.455*X_{18}*X_{15} + 0.22*X_{17}*X_{18}*X_{15} - 0.644*X_{17}^3 + 1.87*X_{18}^3$
 (5) $X_{22} = 0.0673 + 0.99*X_{21} + 0.394*X_{19} - 0.0556*X_2 -$
 $0.201*X_{21}^2 - 1.86*X_{19}^2 + 0.0374*X_2 + 1.6*X_{21}*X_{19} -$
 $0.0927*X_{21}*X_2 + 0.717*X_{21}*X_{19}*X_2 - 0.217*X_{21}^3 - 2.04*X_{19}^3 -$
 $0.0123*X_2^3$
 (6) $X_{23} = -0.0335 + 0.83*X_{22} + 0.161*X_{20} - 0.0189*X_8 -$
 $0.145*X_{20}^2 + 0.0069*X_8^2 + 0.693*X_{22}*X_{20} + 0.239*X_{22}*X_8 -$
 $0.134*X_{20}*X_8 + 0.541*X_{20}^3$

UNITIZER:

$STATUS = 0.362 + 0.481*X_{23}$

APPENDIX C. IXL RULES

A. NUMBER OF CORRECT PREDICTIONS BY RULE

RULE	1:	44
RULE	2:	37
RULE	3:	40
RULE	4:	43
RULE	5:	39
RULE	6:	41
RULE	7:	44
RULE	8:	39
RULE	9:	36
RULE	10:	41
RULE	11:	44
RULE	12:	41
RULE	13:	45
RULE	14:	39
RULE	15:	36
RULE	16:	42
RULE	17:	35
RULE	18:	42
RULE	19:	45
RULE	20:	39
RULE	21:	36
RULE	22:	42
RULE	23:	45
RULE	24:	35
RULE	25:	38
RULE	26:	41
RULE	27:	35
RULE	28:	37
RULE	29:	40
RULE	30:	36
RULE	31:	43
RULE	32:	39
RULE	33:	37
RULE	34:	39

B. IXL RULES

IXL DISCOVERY MODULE

Minimum Certainty: 85.0
Maximum Error Margin: 10.0
Level of Significance: 60.0
Minimum Group Size: 20
Minimum Group Percentage: 0.0
Maximum Rule Length: 7
Minimum Generality: 5
Maximum Generality: 100
Generality Increments: 100

GOAL 1: "status" = "NOT IN MILITARY"
Number of Goals: 1

CF (entire database): 63.8 %
% Rule 1
CF = 85
"status" = "NOT IN MILITARY"
IF
"civjob" = "GOOD CIVJOB"
AND
"16" <= "entryage" <= "19"
AND
"rank" = "E4"
AND
"ethnic" = "WHITE"
AND
"marital" = "SINGLE NO CHILDREN"
AND
"hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
;
% Margin of Error: 9.3 %
% Applicable percentage of sample: 9.9 %
% Applicable number of records: 67

% Rule 2
CF = 88
"status" = "NOT IN MILITARY"
IF
"civjob" = "GOOD CIVJOB"
AND
"16" <= "entryage" <= "19"
AND
"ethnic" = "WHITE"
AND
"marital" = "SINGLE NO CHILDREN"
AND
"techocc" = "NOT IN TECHNICAL OCCUPATION"

```

AND
  "milbene" = "SATISFIED"
;
% Margin of Error: 9.8 %
% Applicable percentage of sample: 7.5 %
% Applicable number of records: 51

% Rule 3
CF = 87
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "16" <= "entryage" <= "19"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 9.3 %
% Applicable percentage of sample: 9.0 %
% Applicable number of records: 61

% Rule 4
CF = 85
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "16" <= "entryage" <= "19"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 9.2 %
% Applicable percentage of sample: 10.0 %
% Applicable number of records: 68

```

```

% Rule 5
CF = 85
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "16" <= "entryage" <= "19"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 9.2 %
% Applicable percentage of sample: 10.0 %
% Applicable number of records: 68

% Rule 6
CF = 86
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "16" <= "entryage" <= "20"
AND
  "rank" = "E4"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "techocc" = "NOT IN TECHNICAL OCCUPATION"
;
% Margin of Error: 9.4 %
% Applicable percentage of sample: 9.3 %
% Applicable number of records: 63

% Rule 7
CF = 86
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "16" <= "entryage" <= "20"
AND
  "rank" = "E4"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"

```

```

;
% Margin of Error: 8.4 %
% Applicable percentage of sample: 11.5 %
% Applicable number of records: 78

% Rule 8
CF = 88
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "16" <= "entryage" <= "20"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 9.4 %
% Applicable percentage of sample: 8.4 %
% Applicable number of records: 57

% Rule 9
CF = 87
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "16" <= "entryage" <= "20"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
  "milbene" = "SATISFIED"
;
% Margin of Error: 9.4 %
% Applicable percentage of sample: 8.8 %
% Applicable number of records: 60

```

```

% Rule 10
CF = 87
  "status" = "NOT IN MILITARY"
  IF
    "civjob" = "GOOD CIVJOB"
  AND
    "16" <= "entryage" <= "20"
  AND
    "ethnic" = "WHITE"
  AND
    "marital" = "SINGLE NO CHILDREN"
  AND
    "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
  AND
    "millife" = "DISSATISFIED"
  ;
% Margin of Error: 8.6 %
% Applicable percentage of sample: 10.3 %
% Applicable number of records: 70

% Rule 11
CF = 86
  "status" = "NOT IN MILITARY"
  IF
    "civjob" = "GOOD CIVJOB"
  AND
    "16" <= "entryage" <= "20"
  AND
    "ethnic" = "WHITE"
  AND
    "marital" = "SINGLE NO CHILDREN"
  AND
    "millife" = "DISSATISFIED"
  ;
% Margin of Error: 8.2 %
% Applicable percentage of sample: 11.8 %
% Applicable number of records: 80

% Rule 12
CF = 86
  "status" = "NOT IN MILITARY"
  IF
    "civjob" = "GOOD CIVJOB"
  AND
    "16" <= "entryage" <= "21"
  AND
    "rank" = "E4"
  AND
    "ethnic" = "WHITE"
  AND
    "marital" = "SINGLE NO CHILDREN"
  AND
    "techocc" = "NOT IN TECHNICAL OCCUPATION"

```

```

AND
  "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
;
% Margin of Error: 9.4 %
% Applicable percentage of sample: 9.3 %
% Applicable number of records: 63

% Rule 13
CF = 85
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "16" <= "entryage" <= "21"
AND
  "rank" = "E4"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
;
% Margin of Error: 8.3 %
% Applicable percentage of sample: 12.1 %
% Applicable number of records: 82

% Rule 14
CF = 86
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "16" <= "entryage" <= "21"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 9.6 %
% Applicable percentage of sample: 8.7 %
% Applicable number of records: 59

```

```

% Rule 15
CF = 85
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "16" <= "entryage" <= "21"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
  "milbene" = "SATISFIED"
;
% Margin of Error: 9.6 %
% Applicable percentage of sample: 9.1 %
% Applicable number of records: 62

% Rule 16
CF = 86
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "16" <= "entryage" <= "21"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 8.8 %
% Applicable percentage of sample: 10.4 %
% Applicable number of records: 71

% Rule 17
CF = 86
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "16" <= "entryage" <= "21"
AND
  "millife" = "SOMEWHAT SATISFIED"
;

```



```

% Margin of Error: 9.4 %
% Applicable percentage of sample: 9.3 %
% Applicable number of records: 63

% Rule 18
CF = 85
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "17" <= "entryage" <= "21"
AND
  "rank" = "E4"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
  "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
;
% Margin of Error: 9.6 %
% Applicable percentage of sample: 9.1 %
% Applicable number of records: 62

% Rule 19
CF = 85
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "17" <= "entryage" <= "21"
AND
  "rank" = "E4"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
;
% Margin of Error: 8.4 %
% Applicable percentage of sample: 11.8 %
% Applicable number of records: 80

% Rule 20
CF = 88
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJCB"
AND
  "17" <= "entryage" <= "21"

```

```

AND
    "ethnic" = "WHITE"
AND
    "marital" = "SINGLE NO CHILDREN"
AND
    "techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
    "millife" = "DISSATISFIED"
;
% Margin of Error: 9.4 %
% Applicable percentage of sample: 8.4 %
% Applicable number of records: 57

% Rule 21
CF = 87
    "status" = "NOT IN MILITARY"
IF
    "civjob" = "GOOD CIVJOB"
AND
    "17" <= "entryage" <= "21"
AND
    "ethnic" = "WHITE"
AND
    "marital" = "SINGLE NO CHILDREN"
AND
    "techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
    "milbene" = "SATISFIED"
;
% Margin of Error: 9.4 %
% Applicable percentage of sample: 8.8 %
% Applicable number of records: 60

% Rule 22
CF = 87
    "status" = "NOT IN MILITARY"
IF
    "civjob" = "GOOD CIVJOB"
AND
    "17" <= "entryage" <= "21"
AND
    "ethnic" = "WHITE"
AND
    "marital" = "SINGLE NO CHILDREN"
AND
    "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
AND
    "millife" = "DISSATISFIED"
;
% Margin of Error: 8.7 %
% Applicable percentage of sample: 10.1 %
% Applicable number of records: 69

```

```

% Rule 23
CF = 85
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "17" <= "entryage" <= "21"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 8.3 %
% Applicable percentage of sample: 12.1 %
% Applicable number of records: 82

% Rule 24
CF = 85
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "17" <= "entryage" <= "21"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
  "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 9.9 %
% Applicable percentage of sample: 8.8 %
% Applicable number of records: 60

% Rule 25
CF = 85
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "17" <= "entryage" <= "21"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 9.2 %

```

```

% Applicable percentage of sample: 10.0 %
% Applicable number of records: 68

% Rule 26
CF = 85
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "17" <= "entryage" <= "21"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 8.4 %
% Applicable percentage of sample: 11.9 %
% Applicable number of records: 81

% Rule 27
CF = 85
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "17" <= "entryage" <= "21"
AND
  "millife" = "SOMEWHAT SATISFIED"
;
% Margin of Error: 9.7 %
% Applicable percentage of sample: 9.0 %
% Applicable number of records: 61

% Rule 28
CF = 86
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
  "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 9.7 %
% Applicable percentage of sample: 8.5 %

```

```

% Applicable number of records: 58

% Rule 29
CF = 87
"status" = "NOT IN MILITARY"
IF
"civiljob" = "GOOD CIVJOB"
AND
"ethnic" = "WHITE"
AND
"marital" = "SINGLE NO CHILDREN"
AND
"techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
"millife" = "DISSATISFIED"
;
% Margin of Error: 8.9 %
% Applicable percentage of sample: 9.9 %
% Applicable number of records: 67

% Rule 30
CF = 85
"status" = "NOT IN MILITARY"
IF
"civiljob" = "GOOD CIVJOB"
AND
"ethnic" = "WHITE"
AND
"marital" = "SINGLE NO CHILDREN"
AND
"techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
"milbene" = "SATISFIED"
;
% Margin of Error: 9.3 %
% Applicable percentage of sample: 9.9 %
% Applicable number of records: 67

```

```

% Rule 31
CF = 86
  "status" = "NOT IN MILITARY"
IF
  "civjob" = "GOOD CIVJOB"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "hsstatus" = "HIGH SCHOOL DIPLOMA GRAD"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 8.4 %
% Applicable percentage of sample: 11.5 %
% Applicable number of records: 78

% Rule 32
CF = 86
  "status" = "NOT IN MILITARY"
IF
  "16" <= "entryage" <= "19"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 9.6 %
% Applicable percentage of sample: 8.7 %
% Applicable number of records: 59

% Rule 33
CF = 86
  "status" = "NOT IN MILITARY"
IF
  "16" <= "entryage" <= "19"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
  "milbene" = "SATISFIED"
;
% Margin of Error: 9.9 %
% Applicable percentage of sample: 8.4 %
% Applicable number of records: 57

```

```
% Rule 34
CF = 85
  "status" = "NOT IN MILITARY"
IF
  "16" <= "entryage" <= "20"
AND
  "ethnic" = "WHITE"
AND
  "marital" = "SINGLE NO CHILDREN"
AND
  "techocc" = "NOT IN TECHNICAL OCCUPATION"
AND
  "millife" = "DISSATISFIED"
;
% Margin of Error: 9.2 %
% Applicable percentage of sample: 10.0 %
% Applicable number of records: 68
```

Number of Rules found: 34

IXL finished normally

APPENDIX D

EXPANDED VARIABLE AIM MODEL

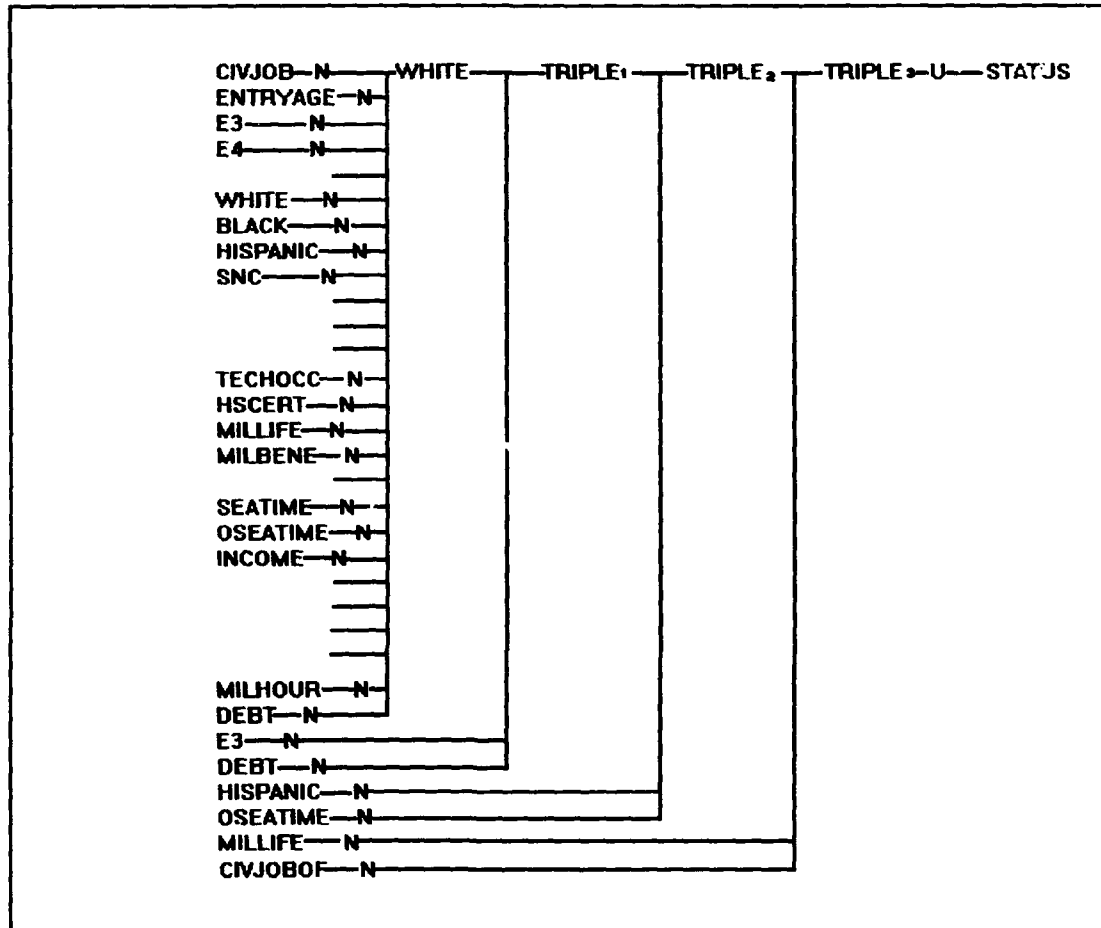


Figure 11: Expanded Variable AIM Network

NORMALIZERS:

$CIVJOB\ X_1 = -2.08 + 2.56 * CIVJOB$
 $ENTRYAGE\ X_2 = -8.64 + 0.448 * ENTRYAGE$
 $E3\ X_3 = -0.511 + 2.46 * E3$
 $E4\ X_4 = -0.794 + 2.05 * E4$
 $WHITE/OTH\ X_6 = -1.97 + 2.48 * WHITE/OTH$
 $BLACK\ X_7 = -0.373 + 3.05 * BLACK$
 $HISPANIC\ X_8 = -0.299 + 3.63 * HISPANIC$
 $SNC\ X_9 = -1.24 + 2.04 * SNC$
 $SWC\ X_{10} = -0.134 + 7.59 * SWC$
 $TECHOCC\ X_{13} = -0.673 + 2.16 * TECHOCC$
 $HSCERT\ X_{14} = -0.439 + 2.71 * HSCERT$
 $MILLIFE\ X_{15} = -0.011 + 1.13 * MILLIFE$

MILBENE $X_{16} = -0.0105 + 1.16 * \text{MILBENE}$
 SEATIME $X_{18} = -1.6 + 0.0587 * \text{SEATIME}$
 OSEATIME $X_{19} = -0.723 + 0.0782 * \text{OSEATIME}$
 INCOME $X_{20} = -1.91 + 0.000136 * \text{INCOME}$
 CIVJOB OF $X_{24} = -1.19 + 2.03 * \text{CIVJOB OF}$
 MILHOUR $X_{25} = -2.81 + 0.00221 * \text{MILHOUR}$
 DEBT $X_{26} = -1.64 + 0.585 * \text{DEBT}$

WHITE:

WHITE $X_{27} = -0.122 * X_1 + 0.0504 * X_2 - 0.126 * X_3 - 0.0504 * X_4 -$
 $0.238 * X_6 - 0.0938 * X_7 - 0.172 * X_8 - 0.132 * X_9 + 0.0499 * X_{13} +$
 $0.0355 * X_{14} + 0.126 * X_{15} + 0.0769 * X_{16} + 0.0533 * X_{18} + 0.0792 * X_{19}$
 $+ 0.103 * X_{20} + 0.0579 * X_{25} + 0.0455 * X_{26}$

TRIPLES:

TRIPLE₁ $X_{28} = 0.174 + 1.05 * X_{27} + 0.324 * X_3 + 0.0954 * X_{26} -$
 $0.348 * X_{27} * X_3 - 0.0869 * X_{27} * X_{26} + 0.166 * X_{27} * X_3 * X_{26} - 0.192 * X_{27}^3$
 $- 0.138 * X_3^3 - 0.0297 * X_{26}^3$
 TRIPLE₂ $X_{29} = 0.992 * X_{28} + 0.0363 * X_{19} + 0.314 * X_{28}^2 -$
 $0.00892 * X_8^2 - 0.0254 * X_{19}^2 + 0.182 * X_{28} * X_8 - 0.226 * X_{28} * X_{19} -$
 $0.0839 * X_8 * X_{19} + 0.134 * X_{28} * X_8 * X_{19} + 0.00566 * X_{19}$
 TRIPLE₃ $X_{30} = 0.964 * X_{29} - 0.0412 * X_{15}^2 + 0.14 * X_{29} * X_{15} -$
 $0.179 * X_{29} * X_{24} + 0.0473 * X_{15} * X_{24}$

UNITIZER:

U $X_{31} = 0.362 + 0.481 * X_{30}$

LIST OF REFERENCES

1. Angell, B. and Murphy, T., "Faster Than a Speeding Neural Network," *AI Expert*, v.7, November 1992.
2. Sands, W.A. and Wilkins, C.A., "A Comparison of Ordinary Least-Squares-Linear Regression and Artificial Neural Network - Back Propagation Models for Personnel Selection Decisions," paper presented at the 1993 Conference on Artificial Neural Networks, San Diego, CA., 3 February 1993.
3. Wiggins, V., Grobman, J., and Looper, L., "Statistical Neural Network Analysis Package (SNNAP)," paper presented at the 1993 Conference on Artificial Neural Networks, San Diego, Ca., 3 February 1993.
4. Marquez, L., Hill, T., Worthley, R., and Remus, W., "Neural Network Models As An Alternative To Regression," working paper, University of Hawaii, Honolulu, Hi., 1991.
5. Hill, T., O'Connor, M., and Remus, W., "Neural Network Models For Time Series Forecasts," working paper, University of Hawaii, Honolulu, Hi., 1992.
6. Weiss, S. and Kapouleas, I., "An Empirical Comparison of Pattern Recognition, Neural Nets, and Machine Learning Classification Methods," *Readings In Machine Learning*, Morgan Kaufmann Publishers, Inc., 1990.
7. Mooney, R. and others, "An Experimental Comparison of Symbolic and Connectionist Learning Algorithms," *Readings In Machine Learning*, Morgan Kaufmann Publishers, Inc., 1990.
8. Naval Research Laboratory Report 9330, *Machine Learning Systems: Part I- Concept Learning from Examples with AQ15 and Related Systems*, by Gordon, D. and Spears, W., 30 September 1991.
9. *AIM User's Manual*, AbTech Corporation, 1992
10. *AIM User's Manual*, AbTech Corporation, 1992.
11. Durkin, J., "Induction... via ID3," *AI Expert*, v.7, April 1992.
12. IntelligenceWare Technical Report, "Discovering Knowledge From Large Databases," by Parsaye, K., and Hansson, O., 1987.

13. *IXL User's Manual*, IntelligenceWare, 1990.

14. Mehay, S., "Post-Service Earnings of Volunteer-Era Veterans: Evidence From the Reserves," working paper, Naval Postgraduate School, Monterey, Ca., May 1992.

INITIAL DISTRIBUTION LIST

	No. Copies
1. Defense Technical Information Center Cameron Station Alexandria VA 22304-6145	2
2. Library, Code 052 Naval Postgraduate School Monterey CA 93943-5002	2
3. Prof. George Thomas, Code AS/TE Dept. of Administrative Science Naval Postgraduate School Monterey, CA 93943-5002	1
4. Prof. Daniel Dolk, Code AS/DK Dept. of Administrative Science Naval Postgraduate School Monterey, CA 93943-5002	1
5. LT. Dennis E. Pytel, Jr. 1308 Sussex Pl. Norfolk, VA. 23508	2
6. Mr. Keith Drake AbTech Corporation 700 Harris Street Charlottesville, VA. 22903	1